

**ΕΡΓΑΛΕΙΑ ΥΠΟΣΤΗΡΙΞΗΣ ΣΗΜΑΣΙΟΛΟΓΙΑΣ
ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΕΡΩΤΗΣΕΩΝ
ΣΕ ΨΗΦΙΑΚΟ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟ ΟΙΚΟΣΥΣΤΗΜΑ**

ΙΩΑΝΝΗΣ-ΑΡΙΣΤΟΤΕΛΗΣ ΚΟΤΟΠΟΥΛΟΣ

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΧΑΝΙΑ 2011

**ΑΝΑΠΤΥΞΗ ΣΥΣΤΗΜΑΤΟΣ ΚΑΙ ΜΕΘΟΔΟΛΟΓΙΑΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΚΑΙ
ΑΝΑΛΥΣΗΣ ΕΡΩΤΗΣΕΩΝ
ΣΕ ΠΕΡΙΒΑΛΛΟΝ ΣΗΜΑΣΙΟΛΟΓΙΚΟΥ ΔΙΚΤΥΟΥ**

ΙΩΑΝΝΗΣ-ΑΡΙΣΤΟΤΕΛΗΣ ΚΟΤΟΠΟΥΛΟΣ

Μια διατριβή που παρουσιάστηκε στο Πολυτεχνείο Κρήτης
σε εκπλήρωση των υποχρεώσεων για το
Μεταπτυχιακό Δίπλωμα Ειδίκευσης του τμήματος των
Ηλεκτρονικών Μηχανικών & Μηχανικών Υπολογιστών

Χανιά Κρήτης, Ελλάδα
© Ιωάννης-Αριστοτέλης Κοτόπουλος 2011

Δηλώνω ότι είμαι ο μοναδικός συγγραφέας αυτής της διατριβής. Εξουσιοδοτώ το Πολυτεχνείο Κρήτης να δανείζει την διατριβή σε άλλους οργανισμούς ή μεμονωμένα άτομα για ερευνητικούς σκοπούς. Επίσης εξουσιοδοτώ το Πολυτεχνείο Κρήτης να αναπαράγει αυτή τη διατριβή με φωτοτυπίες ή άλλους τρόπους, ολόκληρη ή μέρος της, μετά από αίτηση άλλων οργανισμών ή μεμονωμένων ατόμων για ερευνητικούς σκοπούς.

Το Πολυτεχνείο Κρήτης ζητά τις υπογραφές όλων των προσώπων που χρησιμοποιούν ή αντιγράφουν αυτή τη διατριβή. Παρακαλώ υπογράψτε παρακάτω αναγράφοντας διεύθυνση και ημερομηνία.

Περίληψη

Η έκρηξη του λεγόμενου σημασιολογικού ιστού την τελευταία δεκαετία εκπορεύτηκε από την ανάγκη αναγνώρισης της σημασίας σε σχέση με την απλή περιγραφή των αντικειμένων. Μια γενικότερη στροφή στη σημασιολογία, άνοιξε πολλές συζητήσεις, ανάγκες και δημιούργησε νέες προκλήσεις σε όλους τους τομείς έρευνας που σχετίζονταν με την εκμετάλλευση του παγκόσμιου ιστού. Παράλληλα η ίδια η ανάπτυξη του δικτύου έδωσε τη δυνατότητα σύνδεσης όλο και περισσότερων χρηστών μεταξύ τους, εξυπηρετώντας διαφορετικές ανάγκες κάθε φορά και διευρύνοντας τους φυσικούς τρόπους σύνδεσης μεταξύ τους. Στην παρούσα εργασία αναλάβαμε τη δημιουργία ενός κατανεμημένου συστήματος διαχείρισης «γνώσης» μεταξύ μικρομεσαίων επιχειρήσεων με στόχο τη διαφήμιση των υπηρεσιών τους και την διευκόλυνση των συνεργασιών μεταξύ τους.

Για τις ανάγκες του συστήματος ασχοληθήκαμε αρχικά με τη δημιουργία του δικτύου σε φυσικό επίπεδο ενώ βασικός στόχος μας ήταν η κατασκευή και διαχείριση ενός υπερκείμενου δικτύου με βάση τη σημασιολογία. Τα κύρια προβλήματα που είχαμε να αντιμετωπίσουμε ήταν αυτά της ετερογένειας των δεδομένων καθώς και η απαίτηση για αποδοτική ανάκτησή τους από το δίκτυο. Τα εργαλεία που χρησιμοποιήσαμε ήταν οι οντολογίες και μια σειρά από πρότυπες τεχνολογίες για τη διαχείρισή τους αλλά και την εκμετάλλευση της σημασιολογίας τους. Σημαντικό ρόλο έπαιξε η αρχιτεκτονική του συστήματος ενώ απαιτήθηκε έρευνα στο κομμάτι της «έξυπνης» ανάκτησης των δεδομένων από τους κόμβους του ομότιμου δικτύου. Για το σκοπό αυτό δημιουργήθηκε ένα υποσύστημα διαχείρισης σύνθετων ερωτήσεων απευθυνόμενων στο δίκτυο. Η εργασία ολοκληρώθηκε με έρευνα στο θέμα της ανάλυσης των ερωτήσεων και της εκτίμησης του κόστους τους ώστε να αντιμετωπιστεί το θέμα της αποδοτικής επεξεργασίας και προώθησής τους.

Αφιέρωση
στους γονείς μου

Ευχαριστήρια

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Σ. Χριστοδουλάκη για την καθοδήγησή του και τη συνεργασία μας όλα αυτά τα χρόνια. Οι ιδέες του και οι εμπειρία του παίζανε το σημαντικότερο ρόλο ενώ δεν πρέπει να παραλείψω να αναγνωρίσω το έργο του στο εργαστήριο Διανεμημένων Πληροφοριακών Συστημάτων και Εφαρμογών και στο Πολυτεχνείο Κρήτης, αποτελώντας πάντα το φωτεινό παράδειγμα για όλους μας.

Θα ήθελα επίσης να ευχαριστήσω τους καθηγητές κ.κ. Πετράκη Ευριπίδη και Σαμολαδά Βασίλη για το χρόνο που διέθεσαν για την ανάγνωση του κειμένου μου αλλά και τις όποιες συμβουλές και παρατηρήσεις τους τον προηγούμενο καιρό.

Ιδιαίτερα θα ήθελα να ευχαριστήσω το προσωπικό και τους συνεργάτες του εργαστηρίου και συγκεκριμένα τον Καζάση Φώτη για την επίβλεψή του και την άριστη συνεργασία που είχαμε αλλά και τους Παππά Νίκο, Γιολδάση Νεκτάριο, Ανέστη Γιώργο και Κοτόπουλο Γιώργο για τις ώρες που αφιερώσαμε μαζί μαθαίνοντας, ερευνώντας και δουλεύοντας, βγάζοντας τελικά το καλύτερο αποτέλεσμα.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1 – ΣΤΟΧΟΣ ΚΑΙ ΠΛΑΪΣΙΟ ΕΡΓΑΣΙΑΣ.....11

1.1. ΕΙΣΑΓΩΓΗ	11
1.1.1. ΟΝΤΟΛΟΓΙΕΣ	14
1.1.2. ΜΕΤΑ-ΔΕΔΟΜΕΝΑ	15
1.2. ΠΛΑΪΣΙΟ ΕΡΓΑΣΙΑΣ.....	16
1.3. ΣΤΟΧΟΣ ΕΡΓΑΣΙΑΣ ΚΑΙ ΣΥΝΕΙΣΦΟΡΑ	18
1.4. ΤΟ ΕΡΕΥΝΗΤΙΚΟ ΠΡΟΓΡΑΜΜΑ DIGITAL BUSINESS ECOSYSTEM.....	19
1.5. ΣΧΕΤΙΚΗ ΕΡΓΑΣΙΑ ΚΑΙ ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ	21
1.5.1. ΔΙΚΤΥΑ ΟΜΟΤΙΜΩΝ ΚΟΜΒΩΝ ΚΑΙ ΜΕΘΟΔΟΛΟΓΙΕΣ ΕΠΙΛΟΓΗΣ ΚΟΜΒΩΝ	22
1.5.2. ΕΥΡΕΤΗΡΙΑ ΣΕ ΔΙΚΤΥΑ ΟΜΟΤΙΜΩΝ ΚΟΜΒΩΝ ΚΑΙ ΣΗΜΑΣΙΟΛΟΓΙΚΑ ΥΠΕΡΚΕΙΜΕΝΑ ΔΙΚΤΥΑ.....	33
1.5.3. ΔΡΟΜΟΛΟΓΗΣΗ ΕΡΩΤΗΣΕΩΝ	35
1.5.4. ΕΥΡΕΤΗΡΙΑ ΚΑΙ ΣΥΝΟΨΕΙΣ ΠΑΝΩ ΣΕ XML ΔΕΔΟΜΕΝΑ	36
1.5.5. ΕΚΤΙΜΗΣΗ ΕΠΙΛΕΚΤΙΚΟΤΗΤΑΣ	37
1.5.6. ΕΚΤΙΜΗΣΗ ΚΟΣΤΟΥΣ	40
1.5.7. ΥΠΟΛΟΓΙΣΜΟΣ ΕΠΙΛΕΚΤΙΚΟΤΗΤΑΣ ΣΕ ΑΝΤΙΚΕΙΜΕΝΟΣΤΡΕΦΕΙΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ.....	41
1.5.8. ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΟΜΟΙΟΤΗΤΑ ΚΑΙ ΑΝΤΙΣΤΟΙΧΙΣΗ ΟΝΤΟΛΟΓΙΩΝ	41

ΚΕΦΑΛΑΙΟ 2 – ΤΟ ΠΡΟΓΡΑΜΜΑ DBE ΚΑΙ ΤΟ ΤΕΧΝΙΚΟ ΥΠΟΒΑΘΡΟ44

2.1. ΤΟ ΕΡΕΥΝΗΤΙΚΟ ΠΡΟΓΡΑΜΜΑ DBE.....	44
2.1.1. ΕΠΙΣΤΗΜΟΝΙΚΟΙ ΣΤΟΧΟΙ	46
2.1.2. Η ΣΥΝΕΙΣΦΟΡΑ ΣΤΟ DBE	47
2.2. ΤΕΧΝΙΚΟ ΠΛΑΪΣΙΟ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ	48
2.2.1. ΤΟ ΔΙΚΤΥΟ	48
2.2.2. Η ΒΑΣΗ ΓΝΩΣΗΣ	49
2.3. ΠΕΡΙΓΡΑΦΜΑ ΕΡΓΑΣΙΑΣ.....	51
2.3.1. ΦΥΣΙΚΟ ΔΙΚΤΥΟ	52
2.3.2. ΤΟ FADA (FEDERATED AUTONOMOUS DIRECTORY ARCHITECTURE).....	54
2.4. ΤΕΧΝΙΚΟ ΠΛΑΪΣΙΟ ΕΡΓΑΣΙΑΣ.....	56
2.4.1. ΤΟ ΜΟΦ	61
2.5. Η ΓΛΩΣΣΑ ΕΡΩΤΗΣΕΩΝ QML	64
2.5.1. QML ΕΚΦΡΑΣΕΙΣ ΚΑΙ ΣΗΜΑΣΙΟΛΟΓΙΑ.....	70

ΚΕΦΑΛΑΙΟ 3 – ΑΝΑΠΤΥΞΗ ΣΥΣΤΗΜΑΤΟΣ ΚΑΙ ΣΥΝΕΙΣΦΟΡΑ.....72

3.1. ΕΙΣΑΓΩΓΗ	72
3.2. ΑΝΑΠΤΥΞΗ ΟΜΟΤΙΜΟΥ ΔΙΚΤΥΟΥ	73

3.3. Το ODM ΣΑΝ ΓΛΩΣΣΑ ΠΕΡΙΓΡΑΦΗΣ ΟΝΤΟΛΟΓΙΩΝ.....	75
3.3.1. ΜΕΤΑ-ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΚΑΙ ODM	75
3.4. ΒΑΣΗ ΓΝΩΣΗΣ	84
3.5. ΜΟΝΆΔΑ ΠΡΟΣΒΑΣΗΣ ΣΤΗ ΓΝΩΣΗ	87
3.5.1. ΕΠΕΞΕΡΓΑΣΙΑ ΕΡΩΤΗΣΗΣ.....	90
3.6. Η ΜΗΧΑΝΗ ΕΚΤΕΛΕΣΗΣ.....	91
3.6.1. ΠΑΡΑΓΩΓΗ ΤΕΛΙΚΗΣ ΕΡΩΤΗΣΗΣ: Η ΠΕΡΙΠΤΩΣΗ ΤΗΣ ΣΧΕΣΙΑΚΗΣ ΒΆΣΗΣ ΔΕΔΟΜΕΝΩΝ	92
3.6.2. ΠΑΡΑΓΩΓΗ ΤΕΛΙΚΗΣ ΕΡΩΤΗΣΗΣ: Η ΠΕΡΙΠΤΩΣΗ ΤΗΣ XML ΒΆΣΗΣ ΔΕΔΟΜΕΝΩΝ	97
3.7. ΥΠΟΣΥΣΤΗΜΑ ΣΥΣΤΆΣΕΩΝ ΚΑΙ ΕΠΕΚΤΑΣΗ ΕΡΩΤΗΣΕΩΝ	101
3.7.1. ΥΠΟΣΥΣΤΗΜΑ ΣΥΣΤΆΣΕΩΝ ΚΑΙ ΠΡΟΦΪΛ ΧΡΗΣΤΩΝ	101
3.8. ΠΡΩΨΘΗΣΗ ΕΡΩΤΗΣΕΩΝ.....	104

ΚΕΦΆΛΑΙΟ 4 – ΥΠΟΛΟΓΙΣΜΌΣ ΕΠΙΛΕΚΤΙΚΌΤΗΤΑΣ ΚΑΙ ΧΡΉΣΗ ΣΗΜΑΣΙΟΛΟΓΙΚΩΝ ΣΥΣΧΕΤΊΣΕΩΝ

4.1. ΕΙΣΑΓΩΓΗ: ΕΠΙΛΕΚΤΙΚΌΤΗΤΑ ΚΑΙ ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΣΧΕΤΙΚΌΤΗΤΑ ΣΕ ODM ΟΝΤΟΛΟΓΊΕΣ.....	107
4.1.1. ΤΟ ΠΛΆΙΣΙΟ ΕΡΓΑΣΊΑΣ	108
4.1.2. ODM ΚΑΙ ΣΗΜΑΣΙΟΛΟΓΊΑ.....	109
4.2. ΕΠΙΛΕΚΤΙΚΌΤΗΤΑ	110
4.2.1. Η ΑΡΧΗ ΤΗΣ ΑΝΕΞΑΡΤΗΣΊΑΣ.....	113
4.2.2. ΜΕΘΟΔΟΛΟΓΊΑ	114
4.3. ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΣΧΕΤΙΚΌΤΗΤΑ	120
4.3.1. ΘΕΩΡΊΑ ΠΛΗΡΟΦΟΡΊΑΣ ΚΑΙ ΥΠΟΛΟΓΙΣΜΌΣ ΟΜΟΊΟΤΗΤΑΣ.....	121
4.3.2. ODM ΚΑΙ ΥΠΟΛΟΓΙΣΜΌΣ ΣΧΕΤΙΚΌΤΗΤΑΣ.....	125
4.3.3. ΥΠΟΛΟΓΙΣΜΌΣ ΣΗΜΑΣΙΟΛΟΓΙΚΗΣ ΟΜΟΊΟΤΗΤΑΣ.....	126
4.3.4. ΜΕΘΟΔΟΛΟΓΊΑ	128
4.4. ΕΦΑΡΜΟΓΈΣ	136

ΚΕΦΆΛΑΙΟ 5^ο - ΑΝΑΚΕΦΑΛΑΊΩΣΗ ΚΑΙ ΣΥΜΠΕΡΆΣΜΑΤΑ.....

5.1. ΕΙΣΑΓΩΓΗ	138
5.2. ΣΧΕΔΙΑΣΜΌΣ ΚΑΙ ΥΛΟΠΌΙΗΣΗ.....	138
5.2.1. ΤΟ ΔΊΚΤΥΟ	138
5.2.2. Η ΒΆΣΗ ΓΝΩΣΗΣ ΚΑΙ ΤΟ ODM.....	139
5.2.3. ΜΗΧΑΝΙΣΜΟΊ ΑΝΆΚΤΗΣΗΣ	140
5.2.4. ΚΌΣΤΟΣ ΕΡΩΤΗΣΕΩΝ	141

ΒΙΒΛΙΟΓΡΑΦΊΑ.....

ΠΑΡΆΡΤΗΜΑ Α – Η ΓΛΩΣΣΑ ΠΕΡΙΓΡΑΦΗΣ ΟΝΤΟΛΟΓΊΩΝ ODM

Πρόλογος

Η εργασία αυτή ήταν δύσκολο να μπει κάτω από έναν τίτλο γιατί ακολούθησε περισσότερο μια πορεία παρά έναν και μόνο στόχο. Το αντικείμενο αν και ήταν από την αρχή εντοπισμένο διέθετε τα στοιχεία εκείνα που αποτέλεσαν την πρόκληση για την εξυπηρέτηση περισσότερων στόχων με μοναδικό ερωτηματικό τη διάρκεια της εργασίας. Τελικά, μπορεί τα 4 χρόνια να είναι έξω από τα όρια μιας μεταπτυχιακής διατριβής όμως το αποτέλεσμα ήταν πιο πλούσιο και ευρύ και αυτό θα φανεί στις σελίδες που ακολουθούν. Οι νέες τεχνολογίες γύρω από την περιγραφή της «γνώσης» και τη διαχείριση των δεδομένων μαζί με τα γνωστά προβλήματα της ανταλλαγής δεδομένων μεταξύ διαφορετικών πηγών από τη μια πλευρά και της ανάκτησης πληροφοριών από την άλλη, αποτελούν σίγουρα ένα ευρύ πεδίο στο οποίο είχα την ευκαιρία να εντρυφήσω και να ειδικευτώ. Η πληρότητα της προσέγγισης και η ανάγκη κατάρτισης σε περισσότερα γειτνιάζοντα πεδία, αποτελούν και την καλύτερη απόδειξη της ειδίκευσης στην ευρύτερη περιοχή με την οποία ασχολήθηκα. Εκμεταλλεύτηκα την ελευθερία που δίνει ένας πρόλογος για να δώσω μια προσωπική οπτική σε ότι θα ακολουθήσει, προσπαθώντας κυρίως να βοηθήσω τη κατανόηση της διάρθρωσης του κειμένου αλλά και της εργασίας συνολικά, που πρέπει να υπηρετεί και να συνεισφέρει σε κάποια επιστημονική περιοχή.

Κεφάλαιο 1 – Στόχος και Πλαίσιο Εργασίας

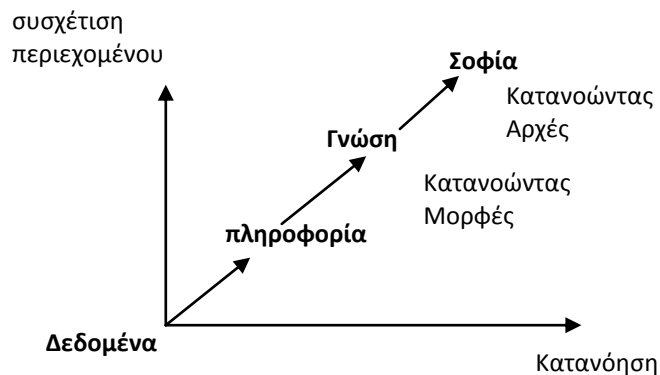
1.1. Εισαγωγή

Η ραγδαία διάδοση του διαδικτύου αποτελεί σήμερα ένα φαινόμενο του οποίου πλέον το πιο ενδιαφέρον στοιχείο δεν είναι ο ρυθμός, αλλά οι τρόποι με τους οποίους επηρεάζει τόσο την καθημερινότητά μας όσο και τις διαφορετικές εκφάνσεις της οικονομικής και κοινωνικής μας εξέλιξης. Παράλληλα η εξέλιξη της τεχνολογίας επέτρεψε τη δημιουργία και αποθήκευση τεράστιου όγκου δεδομένων, στα οποία μπορούσε κανείς, μέσω του διαδικτύου, να έχει και απομακρυσμένη πρόσβαση. Ο όρος δεδομένα δεν εισήχθη ξαφνικά αφού από την αρχή αναφερόταν σε «στοιχεία ή γεγονότα τα οποία συλλέγονται σαν αποτέλεσμα παρατήρησης, εμπειρίας ή πειραματισμού» και από τη στιγμή της συλλογής τους και μετά θεωρούνται γνωστά (δλδ δεδομένα). Μιλάμε δηλαδή για ενέργειες που χαρακτήριζαν από την αρχή τον άνθρωπο και άρα και τα πρώτα δεδομένα πρέπει κανείς να τα αναζητήσει πολύ παλιά. Παρ' όλα αυτά μόνο η οργανωμένη συλλογή και καταγραφή δεδομένων θα μας επιτρέψει τη χρήση του όρου. Τα δεδομένα αποτελούν και την πιο αφηρημένη έννοια στην αλυσίδα περιγραφής της γνώσης με ενδιάμεση έννοια την «πληροφορία». Οι πληροφορίες αποτελούν σήμερα όσο ποτέ άλλοτε, το κύριο αντικείμενο των δραστηριοτήτων του ανθρώπου ενώ η επεξεργασία και εκμετάλλευσή τους οδηγεί στα διάφορα είδη και επίπεδα γνώσης που τον χαρακτηρίζουν. Η εξάπλωση και εξέλιξη του διαδικτύου και της τεχνολογίας, ήταν αυτή που οδήγησε ουσιαστικά στην έκρηξη των διαθέσιμων δεδομένων. Από εκείνο το σημείο και μετά η ανάγκη για επεξεργασία, και διαχείριση των δεδομένων ήταν αυτή που εξ' ορισμού οδήγησε στην ανάγκη για διαχείριση των πληροφοριών και της γνώσης. Οι ανάγκες αυτές δεν ήταν καινούριες, βρέθηκαν όμως σε νέο περιβάλλον και έπρεπε να αντιμετωπιστούν με διαφορετικά μέσα.

Πριν περισσότερο από 5000 χρόνια, στην Μεσοποταμία, λαοί όπως οι Σουμέριοι άρχισαν να δυσκολεύονται με τα χιλιάδες πήλινα πινάκια που είχαν συσσωρευτεί για την καταγραφή νομικών συμβάσεων, φορολογικών εκτιμήσεων, αρχείων πωλήσεων και νόμων. Η λύση ήταν η δημιουργία μιας βιβλιοθήκης, γεγονός που σηματοδοτεί την απαρχή του πρώτου οργανωμένου θεσμού «Διαχείρισης της

Γνώσης». Ωστόσο η πρόσβαση στις πληροφορίες που περιέχονταν εκεί περιοριζόταν κυρίως σε πολιτικούς, θρησκευτικούς ηγέτες και άτομα τα οποία είτε γνώριζαν τη γλώσσα στην οποία ήταν γραμμένα τα πινάκια είτε είχαν τη δυνατότητα να αναθέσουν σε κάποιον τη μετάφρασή τους. Η κατάσταση άλλαξε ουσιαστικά για το ευρύ κοινό με την εφεύρεση του τυπογραφικού πιεστηρίου και της τυπογραφίας, αφού πλέον τα έγγραφα κατά τη διάρκεια της Αναγέννησης (Bergeron, 2003) τυπώνονταν σε μια κοινή γλώσσα. Από τότε και μέχρι σήμερα, η ανάγκη για καταγραφή δεδομένων, δημιουργία και ανταλλαγή πληροφοριών και γνώσης, ακολούθησαν τον άνθρωπο σε όλη του την πορεία παίζοντας πάντα καθοριστικό παράγοντα στην εξέλιξή του. Το ίδιο ισχύει και για την ανάγκη χρήσης κοινών κωδίκων επικοινωνίας και ανταλλαγής πληροφοριών. Η «Διαχείριση της Γνώσης» έχει πλέον αποκτήσει νέους ορισμούς καθώς έχουν μεταβληθεί ριζικά τα μέσα και οι τεχνολογίες αποθήκευσης και επεξεργασίας των δεδομένων. Τελικά, οι πληροφορίες σήμερα έχουν αναχθεί σε έναν πολύτιμο άυλο πόρο ενώ οι μέθοδοι για τη διαχείριση της γνώσης καθώς και όλες οι σχετικές διαδικασίες και τεχνολογίες, έχουν αποκτήσει κεφαλαιώδη σημασία.

Μέχρι αυτό το σημείο έχουμε χρησιμοποιήσει εκ περιτροπής τους όρους «δεδομένα», «πληροφορία» και «γνώση» χωρίς να δώσουμε κάποιον ιδιαίτερο ορισμό αν και αναφερόμαστε σε συγκεκριμένο περιβάλλον. Η αλήθεια είναι ότι η επέλαση της ψηφιακής εποχής έχει απλωθεί σε τόσους τομείς της ανθρώπινης δραστηριότητας ώστε οι ίδιες αυτές οι λέξεις να αποκτήσουν μια πολυσημία, η οποία κάνει δύσκολη οποιαδήποτε προσπάθεια ορισμού τους. Σημαντικότερη και πιο χρήσιμη από τον ίδιο τον ορισμό τους, είναι ίσως η διαδικασία μετάβασης από τα δεδομένα στις πληροφορίες και από τις πληροφορίες στις γνώσεις. Σαν περιγραφή αυτής της διαδικασίας παραθέτουμε το διάγραμμα του σχήματος 1. Σύμφωνα με το διάγραμμα αλλά και τους [(Bellinger, 2004) , (Ackoff, 1989)] η μετάβαση από το ένα επίπεδο στο άλλο σχετίζεται με την απόδοση στα στοιχεία κάθε επιπέδου σημασίας μέσω της κατανόησης και της συσχέτισης τους. Γενικότερα οι λέξεις «κατανόηση», «συσχέτιση» και «σημασία» είναι οι λέξεις κλειδιά στη διαδικασία «Διαχείρισης της Γνώσης».



Σχήμα 1: Αναπαράσταση των Δεδομένων, της Πληροφορίας, της Γνώσης και της Σοφίας (Bellinger, 2004)

Τα πρώτα εργαλεία «Διαχείρισης Γνώσης» που αναπτύχθηκαν κυρίως κατά τη διάρκεια της δεκαετίας του '90, ήταν επηρεασμένα από την (φιλοσοφική) αντίληψη ότι η γνώση αποτελεί αυτοτελές δημιούργημα(με φυσική υπόσταση) το οποίο μπορεί να αποθανατιστεί και να αναπαραχθεί μέσω της τεχνολογίας και το οποίο είναι κατ' ουσίαν, αληθινό (Nonaka & Takeuchi, 1995). Τα περισσότερα εργαλεία προσπαθούσαν να διαχειριστούν τη γνώση ως αντικείμενο και όχι ως στοιχείο ριζωμένο στην ανθρώπινη νόηση και συμπεριφορά. Στην πραγματικότητα όμως η φύση της γνώσης είναι πολύ πιο σύνθετη από τα δεδομένα ή τις πληροφορίες γιατί έχει έντονο το κοινωνικό στοιχείο και απαιτεί την ενεργό συμμετοχή των ανθρώπων. Τα παραπάνω αποκτήσανε τη σημασία τους την επόμενη δεκαετία οπότε και η συζήτηση συμπεριέλαβε τις έννοιες της προηγούμενης παραγράφου: «κατανόηση», «σημασία», «συσχέτιση». Αυτές είναι και οι έννοιες που μας απασχόλησαν στην εργασία αυτή. Έχοντας στο μυαλό μας τα μέσα που χρησιμοποιούνται σήμερα για τη διαχείριση της γνώσης, θα μιλήσουμε για «κατανόηση» αναφερόμενοι στις δυνατότητες κατανόησης των δεδομένων από τους υπολογιστές και εν γένει τις μηχανές. Επίσης θα μιλήσουμε για «σημασία» αναφερόμενοι στη σημασιολογία των δεδομένων και τέλος θα μιλήσουμε για «συσχέτιση» αναφερόμενοι στις φυσικές και σημασιολογικές συσχετίσεις που ενυπάρχουν στα δεδομένα.

Δύο από τους βασικότερους όρους για την περιγραφή της σημασίας και της συσχέτισης των δεδομένων είναι οι όροι «οντολογία» και «μεταδεδομένα». Πάνω σε αυτούς στηρίζονται και οι σύγχρονες περιοχές της επιστήμης που ασχολούνται με την επεξεργασία, διαχείριση και κατανόηση των δεδομένων. Θα κλείσουμε αυτό το

πρώτο μέρος της εισαγωγής με την ανάλυση των δύο αυτών όρων που θα χρησιμοποιήσουμε και σε όλο το υπόλοιπο κείμενο.

1.1.1. Οντολογίες

Ο όρος οντολογία έχει τις ρίζες του στην αρχαία ελληνική φιλοσοφία όπου πρώτος ο Αριστοτέλης τον χρησιμοποίησε ασχολούμενος με την επιστήμη της ύπαρξης. Τότε η Οντολογία¹ προσπαθούσε να απαντήσει στην ερώτηση «τι είναι το on» ή καλύτερα, «ποια είναι τα κοινά χαρακτηριστικά μεταξύ των όντων». Σήμερα ο όρος έχει αποκτήσει διαφορετικό περιεχόμενο ενώ οι ορισμοί του αλλάζουν ανάλογα με το περιβάλλον στο οποίο αναφέρονται. Στα πεδία που μας ενδιαφέρουν δύο από τους πιο κοινούς ορισμούς είναι οι εξής (Gruber T. G., 1993):

- Μια οντολογία είναι ένας σαφής εννοιολογικός προσδιορισμός
- Μια οντολογία αποτελεί μια διαμοιρασμένη και κοινή κατανόηση κάποιου πεδίου η οποία μπορεί να ανταλλάγεί μεταξύ ανθρώπων και συστημάτων εφαρμογών

Με λίγο πιο απλά λόγια από τον ίδιο τον Thomas Gruber, τα παραπάνω σημαίνουν ότι μια οντολογία αποτελεί την περιγραφή των εννοιών(ή των οντοτήτων) καθώς και των σχέσεων μεταξύ τους που μπορεί να υπάρχουν σε έναν πράκτορα ή σε μια κοινότητα από πράκτορες. Με τα παραπάνω γίνεται κατανοητή η σημαντικότητα των οντολογιών στην περιγραφή της γνώσης ενώ επιτρέπουν παράλληλα τον διαμοιρασμό και την επαναχρησιμοποίησή της. Ο κορμός της γνώσης βασίζεται σε έναν εννοιολογικό προσδιορισμό: στα αντικείμενα, τις έννοιες και τις υπόλοιπες οντότητες που υπάρχουν σε μια περιοχή ενδιαφέροντος, καθώς και στις σχέσεις που υπάρχουν μεταξύ τους. Ένας εννοιολογικός προσδιορισμός αποτελεί μια αφηρημένη και απλοποιημένη οπτική του κόσμου τον οποίο θέλουμε να αναπαραστήσουμε για κάποιο σκοπό. Κάθε βάση γνώσης και κάθε σύστημα βασισμένο στη γνώση είναι δεσμευμένο με έναν συγκεκριμένο εννοιολογικό προσδιορισμό. Για αυτά τα συστήματα το τι «υπάρχει» είναι ό,τι μπορεί να αναπαρασταθεί. Όταν η γνώση μιας θεματικής περιοχής αναπαρίσταται δηλωτικά με κάποιον επίσημο τρόπο, το σύνολο των αντικειμένων που μπορούν να

¹ Η Οντολογία με κεφαλαίο χρησιμοποιείται για αναφορά στον κλάδο της φιλοσοφίας σε σχέση με τις οντολογίες στις οποίες θα αναφερόμαστε σαν εργαλεία περιγραφής της γνώσης.

αναπαρασταθούν ονομάζονται το «σύμπαν²» της περιοχής. Αυτό το σύνολο των αντικειμένων και οι περιγραφόμενες μεταξύ τους σχέσεις, αντικατοπτρίζονται στο λεξιλόγιο με το οποίο το σύστημα αναπαριστά τη γνώση. Με αυτόν τον τρόπο η οντολογία ενός συστήματος περιγράφεται με τον ορισμό του συνόλου των όρων που χρησιμοποιεί το σύστημα για την αναπαράσταση της γνώσης. Σε μια τέτοια οντολογία, οι ορισμοί συσχετίζουν τα ονόματα των οντοτήτων με κείμενο κατανοητό από τους ανθρώπους, που περιγράφει τι σημαίνουν τα ονόματα καθώς και αξιώματα που περιορίζουν τη μετάφραση και τη σωστή χρήση των όρων της οντολογίας. Πιο επίσημα θα μπορούσε κανείς να πει ότι μια οντολογία είναι μια δήλωση μιας λογικής θεωρίας (Gruber T. R., 2003) .

Υπάρχουν πολλοί διαφορετικοί ορισμοί για το τι είναι οντολογία σήμερα. Αυτό που δείξαμε με τα παραπάνω είναι τις ανάγκες που εξυπηρετούν οι οντολογίες καθώς και τη χρησιμότητά τους για την περιγραφή και τελικά για τη διαχείριση της γνώσης. Οι οντολογίες αποτελούν ένα βασικό εργαλείο τόσο για την αποτύπωση της σημασιολογίας και των συσχετίσεων μεταξύ των εννοιών μιας θεματικής περιοχής, όσο και για την κατανόησή τους, όπως την είδαμε στην εισαγωγή(κατανόηση-συσχέτιση-σημασία). Επίσης θα πρέπει να τονίσουμε τη χρησιμότητα των οντολογιών στην επίτευξη της διαλειτουργικότητας μεταξύ διαφορετικών συστημάτων καθώς και στο διαμοιρασμό της γνώσης.

1.1.2. Μετα-δεδομένα

Η Οντολογία για την οποία μιλήσαμε στην προηγούμενη παράγραφο, αποτελεί βασικό κλάδο της μεταφυσικής. Η μεταφυσική είναι ο κλάδος της φιλοσοφίας που προσπαθεί να ερμηνεύσει τη φύση της πραγματικότητας του όντος και του κόσμου και το όνομά της αντλεί την προέλευσή του από το έργο «Μετά τα φυσικά» του Αριστοτέλη. Πρόκειται για την πρώτη φορά που το «μετα» χρησιμοποιήθηκε ως πρόθημα μιας λέξης και από τότε χρησιμοποιείται μπροστά από ουσιαστικά με τις εξής διαφορετικές σημασίες ανάλογα με την περίπτωση: μεταξύ, με, μέσω, μετά κ.ά. Στην επιστημολογία που είναι ο κατεξοχήν κλάδος της φιλοσοφίας που ασχολείται με τη θεωρία της γνώσης, η λέξη «μετα» χρησιμοποιείται με τη σημασία του «περι» (τον εαυτό του, σχετικό με τον εαυτό του). Με το ίδιο τρόπο και στην περίπτωση

² Είναι χαρακτηριστική η έκφραση “Universe of discourse”

μας τα μεταδεδομένα είναι δεδομένα σχετικά με τα δεδομένα ή αλλιώς δεδομένα που περιγράφουν τα δεδομένα. Ο όρος απέκτησε μεγαλύτερη αξία μετά την στροφή που έγινε προς τη σημασιολογία. Το να αποδώσει κανείς σημασία στα δεδομένα αναφέρεται ουσιαστικά στην περιγραφή των δεδομένων. Με τον ίδιο τρόπο, σήμερα, τα μεταδεδομένα αναφέρονται κατά κύριο λόγο³ στη σημασιολογία των δεδομένων. Κάτι επίσης σημαντικό είναι ότι προκειμένου να υπάρχει μια περιγραφή κάποιων δεδομένων πρακτικά χρειάζεται ένας πιο «αφηρημένος» προσδιορισμός τους. Αυτή η αφαίρεση καθώς και τα διαφορετικά επίπεδα αφαίρεσης που μπορεί να χρησιμοποιήσει κανείς, είναι σήμερα μια τυπική μεθοδολογία για την περιγραφή της γνώσης. Παράλληλα προκύψαν και μια σειρά από άλλοι όροι τους οποίους θα συναντήσουμε και στην συνέχεια του κειμένου και στηρίζονται ακριβώς στην ίδια ερμηνεία. Ένας από αυτούς είναι ο όρος μεταμοντέλο⁴. Σύμφωνα με τα παραπάνω ένα μεταμοντέλο χρησιμοποιείται για την περιγραφή(την μοντελοποίηση σε αυτή τη περίπτωση) ενός άλλου μοντέλου. Ένα μεταμοντέλο επίσης βρίσκεται σε κάποιο πιο υψηλό επίπεδο αφαίρεσης σε σχέση με το μοντέλο που περιγράφει.

Με όσα είδαμε σε αυτή τη παράγραφο ολοκληρώσαμε το γενικό πλαίσιο γύρω από τη φύση του αντικειμένου της παρούσας εργασίας. Επίσης, δώσαμε τους απαραίτητους ορισμούς που θα βοηθήσουν στη συνέχεια την περιγραφή του ίδιου του αντικειμένου της εργασίας. Αναφερθήκαμε στην ανάγκη διαχείρισης της γνώσης καθώς και στις συνθήκες που δημιουργεί η τεράστια συσσώρευση δεδομένων από τη μια πλευρά και οι νέες τάσεις και τεχνολογίες από την άλλη πλευρά. Στην επόμενη παράγραφο θα δούμε πιο ειδικά το θεωρητικό και τεχνικό πλαίσιο της εργασίας καθώς και τους στόχους της.

1.2. Πλαίσιο Εργασίας

Ο τομέας της επιστήμης της πληροφορικής έχει σήμερα να αντιμετωπίσει τις προκλήσεις της εξάπλωσης του διαδικτύου μαζί με τη συνεχώς αυξανόμενη ανάγκη για διαχείριση και ανταλλαγή δεδομένων και πληροφοριών. Οι εξελίξεις στην

³ Αυτό δεν είναι αποκλειστικό αφού πολλές φορές αναφερόμαστε με τον όρο και στην περιγραφή του τύπου των δεδομένων ή και άλλων χαρακτηριστικών τους που δε θα μπορούσαν να μουν όλα κάτω από τον όρο σημασιολογία.

⁴ Κάποιοι από αυτούς τους όρους δεν έχουν ακόμα καθιερωθεί στην ελληνική παρ' όλα αυτά θα τους χρησιμοποιούμε κανονικά σαν μια λέξη βοηθώντας και στην αντιστοίχισή τους με τους γνωστούς αγγλικούς.

επιστήμη των υπολογιστών όχι απλώς δημιούργησαν κάποια νέα δεδομένα αλλά ταυτόχρονα είναι υπεύθυνες για μια σειρά από τάσεις τόσο από την πλευρά των χρηστών όσο και από την πλευρά των ειδικών στους αντίστοιχους τομείς. Το αποτέλεσμα ήταν παραδοσιακά προβλήματα να πρέπει να εξεταστούν από την αρχή και να αντιμετωπιστούν στο νέο περιβάλλον που δημιούργησαν οι νέες τεχνολογίες. Τα θέματα που μας απασχόλησαν περισσότερο σε αυτήν την εργασία ήταν αυτά της δημιουργίας ενός κατακευκτισμένου συστήματος διαχείρισης δεδομένων μεταξύ ενός δικτύου επιχειρήσεων με στόχο τη συνεργασία τους και τη διαφήμιση των υπηρεσιών τους. Επικεντρωθήκαμε στο κομμάτι της ανάκτησης δεδομένων και της επεξεργασίας ερωτήσεων. Το πλαίσιο στο οποίο εργαστήκαμε συμπεριελάμβανε όλες τις σύγχρονες τάσεις και τεχνολογίες κάτω από την απαίτηση δημιουργίας ενός ολοκληρωμένου συστήματος. Το σύστημα αυτό αναπτύχθηκε για τις ανάγκες του ευρωπαϊκού προγράμματος Digital Business Ecosystem(DBE) για το οποίο θα μιλήσουμε στη συνέχεια. Οι περιοχές στις οποίες χρειάστηκε να εμπλακούμε ήταν οι ακόλουθες:

- Η δημιουργία ενός φυσικού και σημασιολογικού δικτύου μεταξύ των πελατών του συστήματος
- Η αποθήκευση και διαχείριση των δεδομένων περιγραφής των πελατών και των υπηρεσιών τους
- Υπηρεσίες Διαδικτύου για την δημοσίευση και παροχή απομακρυσμένων υπηρεσιών με την λειτουργικότητα του συστήματος
- Η ανάκτηση και επεξεργασία των δεδομένων του συστήματος

Μέσα στις παραπάνω περιοχές διατηρήθηκε σαν κορμός της εργασίας το τρίπτυχο της εισαγωγής, κατανόηση-συσχέτιση-σημασιολογία, δίνοντας έμφαση στον τρόπο με τον οποίο θα γινόταν η επεξεργασία των δεδομένων, η εκμετάλλευση των συσχετίσεων μεταξύ των δεδομένων και τέλος, η αποθήκευση και επεξεργασία της σημασιολογίας τους.

Σε αυτό το πλαίσιο η εργασία κινήθηκε τόσο σε επίπεδο υλοποίησης όσο και σε ερευνητικό επίπεδο απαιτώντας συνολική κατάρτιση γύρω από τις συγκεκριμένες τεχνολογίες καθώς και συνολική εικόνα του συστήματος προκαλώντας με αυτόν τον τρόπο περισσότερα ερεθίσματα και έχοντας πολλαπλή συνεισφορά. Τέλος αξίζει να πούμε ότι το πλαίσιο αυτό προδιαγράφηκε από το πρόγραμμα DBE, τα προβλήματα όμως που αντιμετωπίστηκαν και οι μεθοδολογίες που

χρησιμοποιήθηκαν βρίσκουν εφαρμογή σε κάθε σύγχρονο σύστημα που περιέχει μέρος ή όλα από τα χαρακτηριστικά που διέθετε το πλαίσιό μας.

1.3. Στόχος Εργασίας και Συνεισφορά

Στόχος της εργασίας αυτής, όπως είδαμε και παραπάνω, ήταν η δημιουργία και η διαχείριση μιας βάσης γνώσης για την παροχή υπηρεσιών σε ένα δίκτυο πελατών. Πιο συγκεκριμένα, ο στόχος ήταν η επεξεργασία και η ανάλυση ερωτήσεων σε περιβάλλον σημασιολογικού δικτύου. Η δημιουργία ενός δικτύου από κόμβους εκτός από την απαιτούμενη τεχνογνωσία, απαιτεί την αντιμετώπιση μιας σειρά προβλημάτων όπως η ετερογένεια των δεδομένων, η ανταλλαγή δεδομένων μεταξύ των κόμβων του δικτύου, η αποτελεσματική προώθηση και επεξεργασία ερωτήσεων στο δίκτυο και τέλος η ανάκτηση δεδομένων για συγκεκριμένους σκοπούς. Για τους σκοπούς αυτούς στόχος μας ήταν να δοθεί έμφαση στην σημασιολογική περιγραφή και επεξεργασία των δεδομένων κάτι το οποίο ήταν απαραίτητο στάδιο πριν τη επεξεργασία και ανάλυση ερωτήσεων που ήταν και ο τελικός μας στόχος. Με βάση τα παραπάνω αναπτύχθηκε ένα ολοκληρωμένο σύστημα και η συνεισφορά της εργασίας ήταν:

- ❖ Η ανάπτυξη και προσαρμογή ενός ομοτίμου δικτύου
- ❖ Η σημασιολογική περιγραφή των περιεχομένων των κόμβων με βάση επίσημα πρότυπα που εξασφαλίζουν διαλειτουργικότητα και επεκτασιμότητα(MOF, οντολογίες)
- ❖ Η αποθήκευση και διαχείριση των παραπάνω περιγραφών και δεδομένων (έξυπνοι τρόποι διαχείρισης των μοντέλων, μεταμοντέλων και δεδομένων πάνω από XML βάση δεδομένων και η αρχιτεκτονική που ακολουθήθηκε για να υποστηρίξει μια σειρά από υπηρεσίες και επεκτασιμότητα)
- ❖ Μηχανισμοί επεξεργασίας σύνθετων ερωτήσεων για τα παραπάνω τόσο σε τοπικό επίπεδο όσο και σε επίπεδο δικτύου
- ❖ Η ανάλυση και επέκταση των ερωτήσεων ώστε να υποστηριχθούν ασαφείς ερωτήσεις και πολλαπλές οντολογίες
- ❖ Η προώθηση των ερωτήσεων στο δίκτυο και η συλλογή των αποτελεσμάτων μέσω υπηρεσίας συστάσεων
- ❖ Η πρόταση μεθοδολογίας για την εκτίμηση του κόστους των ερωτήσεων

Τα παραπάνω αφορούν σίγουρα πολλά διαφορετικά κομμάτια του συστήματος και θα μπορούσε κανείς να τα διαχωρίσει σε αυτά που περιλαμβάνανε κυρίως την υλοποίηση και την τεχνική υποδομή του συστήματος και αυτά που σχετίζονται με τη σχεδίαση και τη θεωρητική υποστήριξη και εξέλιξη του συστήματος. Η πραγματικότητα είναι ότι στα πλαίσια της παρούσας εργασίας αποκτήθηκε εμπειρία γύρω από όλα τα παραπάνω θέματα και με αυτόν τον τρόπο υπήρχε η συνολική εικόνα αλλά και τα ερεθίσματα ώστε να επεκτείνουμε το ερευνητικό κομμάτι της εργασίας. Τελικά αφού δημιουργήθηκε ένα σύστημα με συγκεκριμένα χαρακτηριστικά, πραγματοποιήθηκαν συγκεκριμένες δοκιμές και προτάσεις ώστε να αντιμετωπιστούν οι προκλήσεις που αναφέρθηκαν στην αρχή της παραγράφου. Έμφαση δόθηκε στην εκτίμηση του κόστους των ερωτήσεων καθώς και στις στρατηγικές προώθησης των ερωτήσεων στο δίκτυο. Πιο αναλυτικά, στο τελευταίο κομμάτι της εργασίας ασχοληθήκαμε με την εκτίμηση του κόστους ερωτήσεων που απευθύνονται σε σημασιολογικά δεδομένα. Το αποτέλεσμα ήταν η πρόταση για μια μεθοδολογία υπολογισμού του κόστους των ερωτήσεων με στόχο τη βελτίωση της δρομολόγησής τους.

1.4. Το ερευνητικό πρόγραμμα Digital Business Ecosystem

Έχει ήδη αναφερθεί ότι η εργασία αυτή έγινε με αφορμή το ερευνητικό πρόγραμμα DBE και μέσα στα πλαίσια και τις απαιτήσεις που είχε ορίσει αυτό. Το πρόγραμμα προέκυψε σαν αποτέλεσμα του μελλοντικού σχεδιασμού για τις Τεχνολογίες της Κοινωνίας της Πληροφορίας (IST) για το έκτο πλαίσιο στήριξης (FP6) της Ευρωπαϊκής Κοινότητας με σκοπό τη συνεισφορά για τις απαιτήσεις της Λισσαβόνας⁵. Είναι κοινώς αποδεκτό ότι πρέπει να επενδύσουμε περισσότερο και καλύτερα στη γνώση και την καινοτομία οι οποίες αποτελούν οδηγούς για την οικονομική ανάπτυξη. Δεδομένης της σημαντικότητας των μικρομεσαίων επιχειρήσεων για την Ευρωπαϊκή οικονομία⁶, είναι κρίσιμη η ανάπτυξη του κλάδου των μικρομεσαίων επιχειρήσεων καθώς και η υιοθέτηση από αυτές όλο και

⁵ Η συνάντηση των αρχηγών κρατών έγινε στη Λισσαβόνα στις 23-24 Μαρτίου 2000 και τα συμπεράσματά της συνοψίζονται στη δέσμευση για «την πιο ανταγωνιστική και δυναμική οικονομία βασισμένη στη γνώση, ικανή για μια οικονομική ανάπτυξη με περισσότερες και καλύτερες δουλειές για μεγαλύτερη κοινωνική μερίδα» μέχρι το 2010.

⁶ Το 99,7% των εταιριών στην Ευρώπη είναι μικρομεσαίες επιχειρήσεις αντιστοιχώντας στο 50% του Ευρωπαϊκού ΑΕΠ. Το 93% έχει λιγότερους από 10 εργαζομένους (www.ueapme.com).

περισσότερο των τεχνολογιών της πληροφορίας και των επικοινωνιών. Η πρόκληση που αντιμετωπίζουμε περιλαμβάνει τους εξής δύο στόχους:

- Να επιτρέψουμε στους Ευρωπαίους παροχείς πληροφορικών και τηλεπικοινωνιακών τεχνολογιών να γίνουν ανταγωνιστικοί στην αγορά λογισμικού
- Να παρέχουμε λύσεις ηλεκτρονικού επιχειρείν ικανές να προσαρμόζονται στις ανάγκες των τοπικών επιχειρήσεων τροφοδοτώντας την υιοθέτηση τεχνολογιών και την οικονομική ανάπτυξη στους τοπικούς κόμβους καινοτομίας.

Με τους στόχους αυτούς αναπτύχθηκε ένα περίπλοκο εννοιολογικό πλαίσιο για να περιγράψει τις αλληλεπιδράσεις μεταξύ των εταιριών, της τεχνολογίας και της γνώσης, εμπνευσμένο από τα βιολογικά οικοσυστήματα. Βασισμένο σε προηγούμενες δουλειές (Moore, 1996) ανέπτυξε κάποια σημεία σχετικά με την επίσημη περιγραφή της γνώσης για και από μικρομεσαίες επιχειρήσεις με σκοπό τη δημιουργία ενός περιβάλλοντος ανοικτού κώδικα που θα μπορεί να υποστηρίξει την συνεχή εξέλιξη των επιχειρηματικών μοντέλων και υπηρεσιών λογισμικού. Πιο συγκεκριμένα ο στόχος ήταν η εφαρμογή βιολογικών μοντέλων αυτό-οργάνωσης και εξέλιξης στο λογισμικό με βάση την υπόθεση ότι μια τέτοια «βιολογική» συμπεριφορά του λογισμικού είναι πιθανό να βελτιστοποιήσει τη λειτουργία του συστήματος.

Σε ένα τέτοιο σύστημα σημαντικός είναι ο ρόλος του διαμοιρασμού της γνώσης. Ο διαμοιρασμός των ιδεών επιτρέπει τη συνεργατική ανάπτυξη των εφαρμογών αλλά και το χτίσιμο συμβατών συμπεριφορών καθώς και μιας κοινής αντίληψης της πραγματικότητας, δημιουργώντας μια κοινή ταυτότητα και ενθαρρύνοντας την συνάντηση σε επίπεδο «κοινότητας». Για τη δημιουργία αυτής της «κοινότητας» βασικό ρόλο παίζει η διασύνδεση όλων αυτών των επιχειρήσεων-κόμβων μεταξύ τους αναδεικνύοντας την ανάγκη για δημιουργία ενός δικτύου ικανού να ανταποκριθεί στις απαιτήσεις. Τελικά έχουμε ένα δίκτυο με όλες τις ενδιαφερόμενες μικρομεσαίες επιχειρήσεις που επιπλέον τους παρέχει τη δυνατότητα επικοινωνίας, ανταλλαγής δεδομένων και γνώσης και τελικά της παροχής υπηρεσιών για την εξυπηρέτηση των στόχων τους.

Αν τα δούμε συνολικά, για το DBE χρειάστηκε:

- να υλοποιηθεί και να εξελιχθεί η υποδομή δημιουργίας ενός δικτύου από την αρχή
- να χρησιμοποιηθούν τα πρότυπα εκείνα και οι τεχνολογίες που θα επέτρεπαν τη δημιουργία μιας βάσης γνώσης διαμοιραζόμενης από το δίκτυο και διαχειρίσιμης από όλους τους μετέχοντες στο δίκτυο
- η παροχή της υποδομής για τη δημιουργία, τη δημοσίευση, την παροχή και την κατανάλωση υπηρεσιών από τις επιχειρήσεις

τα παραπάνω από μόνα τους δημιουργούν φυσικά κάποιες επιπρόσθετες απαιτήσεις οι οποίες οδήγησαν και στους στόχους που είδαμε για την παρούσα εργασία στην προηγούμενη παράγραφο. Αναλυτικότερα για το πρόγραμμα DBE θα μιλήσουμε στο επόμενο κεφάλαιο τόσο από θεωρητικής όσο και από τεχνικής άποψης.

Στο υπόλοιπο του πρώτου κεφαλαίου θα δούμε τη σχετική εργασία με την εργασία μας σε μια προσπάθεια να κάνουμε πιο σαφές τόσο το πλαίσιο στο οποίο κινηθήκαμε όσο και το υπόβαθρο που απαιτήθηκε για να αντιμετωπιστούν τα θέματα που θα δούμε στα επόμενα κεφάλαια.

1.5. Σχετική εργασία και θεωρητικό υπόβαθρο

Όπως είδαμε τόσο στην εισαγωγή όσο και στους στόχους της εργασίας, το πεδίο με το οποίο ασχοληθήκαμε ήταν αρκετά ευρύ και απαιτήσε πολύπλευρη κατάρτιση σε μια σειρά από διαφορετικές περιοχές. Για το λόγο αυτό η σχετική εργασία που παραθέτουμε είναι εκτενής αποτελώντας ταυτόχρονα και μια αναφορά στη σημερινή κατάσταση των αντίστοιχων περιοχών. Οι εργασίες που παρουσιάζονται είναι είτε απλά σχετικές με τα αντικείμενα της εργασίας μας είτε απαιτήθηκαν σαν υπόβαθρο κατά το σχεδιασμό και ανάπτυξη του πλαισίου που τελικά υλοποιήθηκε. Είναι σημαντικό να επαναλάβουμε και σε αυτό το σημείο ότι η σχεδίαση και υλοποίηση του πλαισίου το οποίο επιλέξαμε σε συνδυασμό με το ολοκληρωμένο υπόβαθρο που θα δούμε στη συνέχεια της παραγράφου, ήταν αυτά που αποτέλεσαν τη βάση για περισσότερες ερευνητικές αναζητήσεις και κατευθύνσεις αυξάνοντας τη συνεισφορά της εργασίας. Για μεγαλύτερη ευκολία στην ανάγνωση η παράγραφος είναι χωρισμένη σε θεματικές ενότητες. Στην παράγραφο 1.5.1 γίνεται αναφορά γενικά στα δίκτυα ομότιμων κόμβων ενώ παρουσιάζεται και η σχετική βιβλιογραφία με την επιλογή κόμβων για την προώθηση ερωτήσεων. Στην

παράγραφο 1.5.2 υπάρχουν ποιο εξειδικευμένες εργασίες που ασχολούνται με τη δημιουργία ευρετηρίων χρήσιμα στην επιλογή κόμβων ενώ γίνεται αναφορά και στα σημασιολογικά υπερκείμενα δίκτυα όπου υπάρχουν μερικές σημαντικές πρόσφατες εργασίες. Η δρομολόγηση ερωτήσεων αποτελεί χωριστή παράγραφο (1.5.3) ενώ η επόμενη παράγραφος (1.5.4) ασχολείται επίσης με τη δημιουργία ευρετηρίων αλλά αυτή τη φορά με αντικείμενο XML δεδομένα. Στις επόμενες δύο παραγράφους (1.5.5, 1.5.6) παρουσιάζονται οι σχετικές εργασίες με την εκτίμηση της επιλεκτικότητας για βάσεις ή για συλλογές δεδομένων καθώς και η εκτίμηση του κόστους των ερωτήσεων. Στην παράγραφο 1.5.7 υπάρχει μια σύντομη αναφορά σε σχετικές εργασίες από την περιοχή των αντικειμενοστραφών βάσεων δεδομένων. Στην τελευταία παράγραφο (1.5.8) υπάρχει η βιβλιογραφία που αναφέρεται στη χρήση οντολογιών και συγκεκριμένα στην αντιμετώπιση του θέματος της αντιστοίχισης οντολογιών αλλά και της εξεύρεσης ομοιοτήτων μεταξύ τους.

1.5.1. Δίκτυα ομότιμων κόμβων και μεθοδολογίες επιλογής κόμβων

Σαν πρώτος μέρος αυτής της εργασίας απαιτήθηκε να γίνει ο σχεδιασμός του δικτύου των μικρομεσαίων επιχειρήσεων πάνω στο οποίο ήταν βασισμένο όλο το πρόγραμμα DBE, όπως είδαμε και νωρίτερα. Έπρεπε για αυτό το λόγο να αποφασιστεί τόσο το είδος του δικτύου που θα υιοθετούσαμε όσο και τα χαρακτηριστικά του. Στην παράγραφο αυτή περιγράφεται η τρέχουσα κατάσταση στην περιοχή των δικτύων ομότιμων κόμβων. Στο επόμενο κεφάλαιο με βάση όσα θα δούμε εδώ, περιγράφουμε τα ακριβή χαρακτηριστικά και τους λόγους για τους οποίους επιλέξαμε να αναπτύξουμε ένα δίκτυο της μορφή που θα δούμε.

Κατά τη διάρκεια ζωής των τεχνολογιών πληροφορικής τα συστήματα υπολογιστών επεκτάθηκαν από συγκεντρωτικά σε κατανεμημένα. Τα κατανεμημένα συστήματα αποτελούνται από υπολογιστές εγκατεστημένους σε διαφορετικές περιοχές με τον καθένα να κάνει ανεξάρτητη επεξεργασία δεδομένων. Συνήθως, σε ένα κατανεμημένο σύστημα κάθε υπολογιστής ειδικεύεται σε μια σειρά από λειτουργίες ή κάποιο συγκεκριμένο έργο. Τα πιο γνωστά και τα πρώτα κατανεμημένα συστήματα είναι αυτά της αρχιτεκτονικής πελάτη-εξυπηρετητή. Σε αυτά τα συστήματα το μέρος του εξυπηρετητή αναλαμβάνει να παρέχει κάποιες κοινές λειτουργίες ενώ το μέρος του πελάτη παρέχει απευθείας υπηρεσίες στους

χρήστες. Το μοντέλο αυτό είναι χρήσιμο με την έννοια ότι επιτρέπει το διαχωρισμό μεταξύ εξειδικευμένων κοινών εργασιών από τις ατομικές εργασίες για τους χρήστες.

Τα τελευταία χρόνια μέσα από τα κατανεμημένα συστήματα είδαμε να αναδεικνύεται το υπολογιστικό παράδειγμα μεταξύ δύο κόμβων. Εκεί γίνεται η προσπάθεια να αρθεί η διαφοροποίηση μεταξύ εξυπηρετητών και πελατών με σκοπό τη δημιουργία ενός δυναμικού δικτύου υπολογιστών όπου οι κόμβοι θα μοιράζονται τους πόρους και τις υπηρεσίες και θα αντιμετωπίζονται σαν ομότιμοι μεταξύ τους. Τέτοια συστήματα υιοθετούνται σε πολλές περιπτώσεις σαν λύση στα προβλήματα που προκαλούν οι περιορισμοί των στατικών τοπολογιών, οι αυξημένες ανάγκες διαχείρισης και η διάχυση των πηγών δεδομένων στο διαδίκτυο. Τα δίκτυα ομότιμων κόμβων μπορούν να θεωρηθούν ως μια αρχιτεκτονική χωρίς κάποιον κεντρικό συντονισμό. Κάθε κόμβος είναι αυτόνομος και μπορεί να λειτουργεί ταυτόχρονα σαν πελάτης και εξυπηρετητής ενώ μοιράζεται περιεχόμενο και πηγές σε απευθείας ανταλλαγή με τους άλλους κόμβους.

Η βιβλιογραφία περιέχει πολλές αρχιτεκτονικές προσεγγίσεις που κατηγοριοποιούν τα δίκτυα ομότιμων κόμβων σε δύο βασικές κατηγορίες: τα υβριδικά και τα καθαρά ομότιμα δίκτυα⁷. Στην υβριδική προσέγγιση μερικές εργασίες πραγματοποιούνται από κεντρικά συστήματα ενώ άλλες είναι αποκεντρωμένες μεταξύ των κόμβων. Στα καθαρά ομότιμα δίκτυα δεν υπάρχει διαφοροποίηση μεταξύ των κόμβων. Καθένας γνωρίζει μόνο γειτονικούς του κόμβους οι οποίοι μάλιστα μπορεί να αλλάζουν με τη πάροδο του χρόνου.

Τα καθαρά ομότιμα δίκτυα μπορεί να είναι είτε δομημένα είτε αδόμητα. Αδόμητα ομότιμα δίκτυα σαν το Gnutella (GNU) χρησιμοποιούν την τεχνική κατάκλισης του δικτύου για την ανακάλυψη των κατάλληλων πηγών δεδομένων. Στην κατάκλιση κάθε κόμβος στέλνει την ερώτηση σε όλους του τους γείτονες. Αν ο γείτονας έχει το αποτέλεσμα θα ενημερώσει τον ερωτώντα και μπορεί να του στείλει απευθείας το αποτέλεσμα διαφορετικά θα μειώσει το Χρόνο Ζωής - TTL⁸ και θα συνεχίσει να προωθεί την ερώτηση στους δικούς του γείτονες για όσο το επιτρέπει ο χρόνος ζωής της ερώτησης. Άλλες προσεγγίσεις αδόμητων ομότιμων δικτύων προσπαθούν

⁷ Στο εξής θα αναφερόμαστε με τον όρο ομότιμα δίκτυα στα δίκτυα των οποίων οι κόμβοι θεωρούνται ομότιμοι.

⁸ Αναφερόμαστε στη γνωστή παράμετρο Time To Live (TTL)

να μειώσουν την επεξεργασία σε όσο το δυνατόν λιγότερους κόμβους. Τέτοιες είναι η επαναληπτική εκβάθυνση (iterative deepening (Garcia-Molina & Yang, 2002)) η οποία εισάγει την ιδέα του προοδευτικού TTL, ο «τυχαίος περίπατος» (random walks (Lv, Cao, Cohen, Li, & Shenker, 2002)) που προσπαθεί να περιορίσει την προώθηση ερωτήσεων στους γείτονες με ευριστικό τρόπο και ενημερωμένη αναζήτηση που εκμεταλλεύεται απλά ευρετήρια κατασκευασμένα από τα αποτελέσματα προηγούμενων ερωτήσεων σε συνδυασμό με κατάκλιση (Kalogeraki, Gunopulos, & Zeinalipour-Yaz, 2002), (Menascie & Kanchanapalli, 2002). Η ενημερωμένη αναζήτηση μπορεί να επεκταθεί για να καλύψει την αντιγραφή(replication) περιεχομένου μαζί με την κατασκευή ευρετηρίου από προηγούμενες ερωτήσεις (Lv, Cao, Cohen, Li, & Shenker, 2002). Τα μεγαλύτερα μειονεκτήματα αυτών των λύσεων είναι ο πολύ μεγάλος όγκος των ανταλλασσόμενων μηνυμάτων, ο πλεονασμός στις ερωτήσεις και τα αποτελέσματα και η δυσκολία καθορισμού του βέλτιστου TTL. Στην ίδια κατηγορία προσεγγίσεων η μετάδοση(broadcasting) έχει αποδείξει τη χρησιμότητά της σε μικρά δίκτυα αλλά και σε μεγάλα ομότιμα δίκτυα διαμοιρασμού αρχείων (Oram, 2001). Η ιδέα είναι ότι οι κόμβοι συνεχίσουν να προωθούν μια ερώτηση στους γείτονές τους, μέχρι να βρεθεί ένας ικανός αριθμός απαντήσεων ή μέχρι έναν μέγιστο αριθμό από προωθήσεις⁹. Αυτή η προσέγγιση δεν δουλεύει για μεγαλύτερες κλίμακες και επίσης είναι πιθανό κάποια δεδομένα να μη βρεθούν ποτέ παρ' όλο που υπάρχουν στο δίκτυο, λόγω του μέγιστου αριθμού προωθήσεων. Το μεγάλο πλεονέκτημα της μετάδοσης είναι ότι έχει πολύ μικρό κόστος συντήρησης και εξάρτησης, δηλαδή πρακτικά δεν χρειάζονται καθόλου μηνύματα για να κρατήσουν το δίκτυο «ζωντανό» και επίσης το δίκτυο είναι πολύ ανθεκτικό σε συχνές αλλαγές στην παρουσία των κόμβων. Στις περιπτώσεις που η μετάδοση είναι απαραίτητη, συγκεκριμένες τοπολογίες όπως το Hypercup (Schlosser M. , Sintek, Decker, & Nejd, 2002) εγγυώνται ότι ένας περιορισμένος αριθμός μηνυμάτων και προωθήσεων χρειάζονται μέχρι να γίνει επικοινωνία με όλους τους κόμβους. Επιπλέον δείχνουν ότι το μοντέλο τους μπορεί να γίνει πιο αποδοτικό με χρήση ενός κοινού σημασιολογικού δικτύου για τον προσδιορισμό της οργάνωσης των κόμβων στη τοπολογία γράφου. Πιο συγκεκριμένα, όταν οι κόμβοι περιγράφουν το περιεχόμενό

⁹ Πρόκειται για το πλήθος των γνωστών ως hops

τους με όρους από αυτήν την κοινή δομή δεδομένων μπορούν να ομαδοποιήσουν τους εαυτούς τους με άλλους όμοιους κόμβους. Αυτή η προσέγγιση, βασισμένη σε ένα δομημένο σε υπερ-κύβο υπερκείμενο δίκτυο έχει μεγαλύτερο κόστος συντήρησης και για αυτόν το λόγο είναι και πιο ευαίσθητη στη δυναμική του δικτύου σε σχέση με τις παραδοσιακές προσεγγίσεις μετάδοσης.

Από την άλλη πλευρά τα δομημένα ομότιμα δίκτυα, κατανέμουν τα δεδομένα στους κόμβους σύμφωνα με κάποιον αλγόριθμο και κάθε κόμβος μπορεί να επιλέξει ποια από αυτά θα κρατήσει (Stoica I. , και συν., Chord: A Scalable Peer-to-peer Lookup Protocol for Internet Applications, 2003) , (Ratnasamy S. , Francis, Handley, Karp, & Shenker, 2001) , (Rowstron & Druschel, Pastry: Scalable, distributed object location and routing for largescale peer-to-peer systems, 2001), (Zhao, Kubiatowicz, & Joseph, April 2001). Τα δομημένα ομότιμα δίκτυα χρειάζονται επιπλέον μηχανισμούς αντιγραφής δεδομένων για να επιτύχουν τη διαθεσιμότητα. Η βασική ιδέα είναι ότι οι κόμβοι αλλά και τα δεδομένα έχουν κάποιο μοναδικό αναγνωριστικό και μοναδικά κλειδιά που παράγονται από συναρτήσεις κατακερματισμού. Κάθε κόμβος είναι υπεύθυνος για την αποθήκευση αρχείων που έχουν κλειδί όμοιο με το αναγνωριστικό του κόμβου. Δοθέντος ενός κλειδιού, ένας κόμβος δρομολογεί την ερώτηση προς τον κόμβο με αναγνωριστικό κοντά στο κλειδί.

Η αρχιτεκτονική με τους υπέρ-κόμβους που περιγράφεται αναλυτικά στο (Yang & Garcia-Molina, 2003) είναι μια υβριδική λύση με ενδιαφέροντα χαρακτηριστικά. Σε τέτοιου είδους δίκτυα οι διαφοροποίηση των κόμβων (απλοί, υπερ-) είναι που εισάγεται σε αντίθεση με τα απλά ομότιμα δίκτυα όπου όλοι οι κόμβοι έχουν ισότιμους ρόλους στο δίκτυο. Ένας υπέρ-κόμβος διατηρεί ένα ευρετήριο προς όλους τους απλούς του κόμβους και είναι ικανός να κάνει ερωτήσεις για λογαριασμό τους. Η αρχιτεκτονική επιτρέπει επίσης απευθείας ανταλλαγή μεταξύ οποιονδήποτε κόμβων. Παρόλο που μια τέτοια προσέγγιση υπόσχεται βελτιωμένη απόδοση σε εργασίες ανακάλυψης, υπάρχουν ακόμα αρκετά ανοιχτά θέματα προς έρευνα όπως ποια είναι μια καλή αναλογία απλών κόμβων προς υπέρ-κόμβους, πως θα πρέπει οι υπέρ-κόμβοι να συνδέονται μεταξύ τους (δομημένα ή όχι), πως μπορούν οι μηχανισμοί αναζήτησης να βελτιωθούν με βάση τη σημασιολογία ή τις σχέσεις μεταξύ δεδομένων, πως το σύστημα μπορεί να είναι ποιο αξιόπιστο χρησιμοποιώντας πλεονασμό κ.ά. Παρ' όλα αυτά, αυτές οι τεχνικές δουλεύουν καλά

για διαμοιρασμό αρχείων, χρησιμοποιώντας τις διαφορετικές χωρητικότητες των κόμβων: κόμβοι με μεγαλύτερη ισχύ, μνήμη ή εύρος δικτύου αναλαμβάνουν επιπλέον εργασίες. Για παράδειγμα το KaZaa (Leibowitz, Ripeanu, & Wierzbicki, 2003), επιτρέπει εθελοντικά στους κόμβους να λειτουργούν σαν υπέρ-κόμβοι που διατηρούν πίνακες δρομολόγησης στους οποίους αποθηκεύεται πληροφορία για το περιεχόμενο άλλων κόμβων (συγκριτικά με χρυσό οδηγό). Αυτή η προσέγγιση εισάγει μιας μορφής συγκεντρωτισμού του συστήματος. Παρ' όλο που είναι καλύτερη από την απλή μετάδοση που είδαμε παραπάνω, στην ουσία παραμένει μετάδοση και για αυτό μπορεί να βελτιωθεί από τεχνικές που κάνουν πιο αποδοτική δρομολόγηση και θα τις δούμε παρακάτω. Οι (Nejdl, και συν., 2003) παρουσιάζουν ομότιμα δίκτυα βασισμένα στο σχήμα των δεδομένων και χρησιμοποιούν τοπολογίες υπέρ-κόμβων στις οποίες οι κόμβοι είναι οργανωμένοι σε υπέρ-κύβους. Οι (Löser, Wolpers, Siberski, & Nejdl, 2003) δείχνουν πως αυτή η προσέγγιση βασισμένη στο σχήμα μπορεί να χρησιμοποιηθεί για τη δημιουργία σημασιολογικών υπερκειμένων ομάδων¹⁰ μέσα σε ένα σημασιολογικό ομότιμο δίκτυο με ένα μικρό σύνολο από χαρακτηριστικά μεταδεδομένα για να περιγραφούν τα έγγραφα του δικτύου.

Μια προσέγγιση που δείχνει να μοιάζει με τους υπέρ-κόμβους αλλά με διαφορετικό στόχο, είναι αυτή των μεσιτών. Η κοινότητα των πολύ-πρακτόρων πρότεινε την έννοια του πράκτορα-μεσίτη όπως στο σύστημα InfoSleuth (Bohrer, και συν., 1997), ο οποίος προσπαθεί να αντιστοιχίσει ανάγκες για πληροφόρηση με τις διαθέσιμες πηγές. Στο InfoSleuth, πράκτορες διαφημίζουν τις υπηρεσίες τους στο μεσίτη με χρήση της γλώσσας KQML (Finin, Fritzson, McKay, & McEntire, 1994). Οι μεσίτες απαντάνε στα αιτήματα των πρακτόρων για υπηρεσίες με πληροφορίες σχετικές με τους πράκτορες που στο παρελθόν διαφήμισαν σχετικές υπηρεσίες. Η βιβλιογραφία για τους πράκτορες-μεσίτες έχει ξεκάθαρο προσανατολισμό στην ανακάλυψη υπηρεσιών. Για αυτό το λόγο δεν είναι περίεργο που η προσέγγιση των μεσιτών είναι πολύ δημοφιλής στη βιβλιογραφία για αναζήτηση διαδικτυακών υπηρεσιών οι οποίες περιγράφονται σημασιολογικά (McIlraith, Son, & Zeng, 2001). Ένα στοιχείο το οποίο δεν είναι πολύ σαφές είναι το κατά πόσο αυτή η προσέγγιση είναι ανθεκτική και κατάλληλη για μεγαλύτερες κλίμακες.

¹⁰ Semantic Overlay Clusters

Μια άλλη τεχνική που προέρχεται από την κοινότητα των ομότιμων δικτύων, κάνει χρήση των κατανεμημένων πινάκων κατακερματισμού DHTs (Distributed Hash Tables). Αυτοί βασίζονται στην ιδέα της δρομολόγησης περιεχομένου στον κόμβο του οποίου το αναγνωριστικό είναι πιο κοντά στο μοναδικό αναγνωριστικό του περιεχομένου. Αυτή η τεχνική υποθέτει πως όλοι οι κόμβοι έχουν την ίδια συνάρτηση κατακερματισμού για να αναθέτουν μοναδικά αναγνωριστικά στο περιεχόμενο, το οποίο θα μπορούσε να είναι οτιδήποτε από έγγραφα, μουσική, διευθύνσεις ως λέξεις. Τα συστήματα που κάνουν δρομολόγηση βασισμένα σε DHTs, όπως το Chord (Stoica I. , Morris, Karger, Kaashoek, & Balakrishnan, 2001) και το Pastry (Rowstron & Druschel, Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems, 2001), είναι ανθεκτικά, στις γρήγορες αλλαγές των μελών του δικτύου. Ένα μειονέκτημα των περισσότερων προσεγγίσεων σε αυτή τη κατηγορία είναι ότι έχουν υψηλό κόστος συντήρησης εξαιτίας των συχνών αλλαγών στο υπερκείμενο δίκτυο σαν αποτέλεσμα της συνεχούς προσθαφαίρεσης μελών σε αυτό. Το PGrid (Aberer, P-grid: a self-organizing access structure for p2p information systems, 2001) είναι ένα σύστημα ομότιμου δικτύου βασισμένο σε ένα εικονικά κατανεμημένο δέντρο αναζήτησης, ανάλογα δομημένο σαν κλασικούς πίνακες κατακερματισμού αλλά με έναν αδόμητο τρόπο κατασκευής του DHT υπερκείμενου δικτύου. Ένα μειονέκτημα των καθαρά βασισμένων σε DHT προσεγγίσεων είναι η κατανομή του φόρτου εργασίας. Εξαιτίας του γεγονότος ότι το περιεχόμενο και οι ερωτήσεις ακολουθούν εκθετική κατανομή, μερικοί κόμβοι(υπεύθυνοι για τα πιο δημοφιλή κλειδιά) είναι πιο φορτωμένοι από κάποιους άλλους. Για αυτό το λόγο χρειάζονται στρατηγικές για την κατανομή του φόρτου εργασίας πάνω από ένα DHT δίκτυο, κάτι το οποίο δεν είναι αναγκαίο σε άλλες περιπτώσεις δικτύων. Επίσης ένα DHT δίκτυο συνήθως είναι λιγότερο ανθεκτικό καθώς φυσιολογικά μόνο ένα κόμβος είναι υπεύθυνος για ένα κλειδί και αν αυτός δεν απαντά τότε το αντίστοιχο περιεχόμενο δεν μπορεί να βρεθεί. Οι δουλειά των (Byers, Considine, & Mitzenmacher, 2001) επιβεβαιώνει το πρόβλημα της κατανομής του φόρτου και περιγράφει μια εναλλακτική προσέγγιση εισάγοντας τον πλεονασμό του περιεχομένου στο δίκτυο, ο οποίος βέβαια απαιτεί επιπλέον κόστος συντήρησης (Haase, και συν.).

Τα συστήματα που είδαμε μέχρι στιγμής δεν είναι κατάλληλα για εφαρμογή στις περιπτώσεις που απαιτείται αποδοτικός διαμοιρασμός δεδομένων (όπως ο διαμοιρασμός σημασιολογικών πληροφοριών σε διαφορετικές βάσεις και κόμβους). Στη βιβλιογραφία μπορεί κανείς να βρει πολλά ερευνητικά αποτελέσματα εγκατάστασης βάσεων δεδομένων σε δίκτυο ομότιμων κόμβων (Ng, Ooi, Tan, & Zhou, 2003), (Sahin, Gupta, Agrawal, & El Abbadi, 2002), (Halevy, Ives, Mork, & Tatarinov, May 2003), (Kantere, Kiringa, Mylopoulos, Kementsietsidis, & Arenas, September 2003), (Arenas, Kantere, Kementsietsidis, Kiringa, Miller, & Mylopoulos, 2003). Η ερευνητική πρόκληση σε τέτοια συστήματα που διαμοιράζονται πλούσια σημασιολογικά δεδομένα, είναι η αποδοτική αντιστοίχιση μεταξύ σχημάτων, η αποδοτική εκτέλεση ερωτήσεων μεταξύ των κόμβων και οι μηχανισμοί που εξασφαλίζουν πλεονασμό και συνέπεια.

Πολλά είναι τα συστήματα που αναπτύσσονται στην περιοχή των ομότιμων δικτύων βασισμένων στο σχήμα¹¹. Τα περισσότερα όμως από αυτά είτε είναι συγκεντρωτικά χωρίς να αντιμετωπίζουν το πρόβλημα των ετερογενών σχημάτων σε κάθε κόμβο είτε είναι αποκεντρωμένα, σαν το δίκτυο του DBE. Αυτά τα συστήματα χρησιμοποιούν κεντρικά ευρετήρια όπως το Napster (Napster) ή ακολουθούν πιο ακριβές προσεγγίσεις όπως η κατάκλιση στο Gnutella (GNU). Άλλα συστήματα όπως το Freenet (Clarke, Sandberg, Wiley, & Hong, Freenet: A distributed anonymous information storage and retrieval system, June 2000) προσπαθούν να χρησιμοποιήσουν πιο αποδοτικές προσεγγίσεις ευρετηρίασης, βασισμένες σε κοινές συναρτήσεις κατακερματισμού και αποθήκευση των τοποθεσιών των πρόσφατων εγγράφων. Υπάρχει επίσης ένας αριθμός ερευνητικών συστημάτων {CAN (Ratnasamy S. , Francis, Handley, Karp, & Shenker, August 2001), CHORD (Stoica I. , και συν., Chord: A Scalable Peer-to-peer Lookup Protocol for Internet Applications, February 2003), Tapestry (Zhao, Kubiawicz, & Joseph, April 2001)} τα οποία αναζητούν αποδοτικά, έγγραφα σε ένα δίκτυο ομότιμων κόμβων. Τα συστήματα όμως αυτά επιβάλλουν μια συγκεκριμένη τοπολογία και κάνουν αναζήτηση πάνω στα αναγνωριστικά των εγγράφων παρά στο περιεχόμενό τους, το οποίο και πάλι δεν είναι το ζητούμενό μας. Στο (Yang & Garcia-Molina, 2001) υπάρχει μια αναφορά στα συστήματα ομότιμων κόμβων με κεντρική αναζήτηση

¹¹ Αναφερόμαστε στο σχήμα περιγραφής των δεδομένων

ενώ στο (Rahm & Bernstein, A survey of approaches to automatic schema matching, 2001) περιέχεται μια έρευνα στα συστήματα για αντιστοίχιση σχημάτων.

Τα συστήματα ομότιμων κόμβων για τη διαχείριση της γνώσης προσπαθούν να εκμεταλλευτούν τη σημασιολογία(γλωσσάρια, μεταδεδομένα, ταξινομίες¹² και οντολογίες) προκειμένου να δημιουργήσουν κατάλληλα ευρετήρια για να διευκολύνουν την κατανομή και ανακάλυψη της γνώσης. Σήμερα υπάρχει γενικότερα μια τάση στη κοινότητα προς αυτήν τη κατεύθυνση και υπάρχουν πολλές τρέχουσες έρευνες. Το ιεραρχικό σχήμα σύνοψης που προτείνεται στο (Tao Shen, Shu, & Yu, 2004) προσπαθεί να χτίσει αποδοτικά και συμπιεσμένα ευρετήρια των εγγράφων(χρησιμοποιώντας το γλωσσάρι των όρων) με σκοπό τη σημασιολογική ανακάλυψη της γνώσης σε συστήματα ομότιμων κόμβων. Η δρομολόγηση βασισμένη σε ταξινομίες που προτείνεται στο (Löser A. , Towards Taxonomy based Routing in P2P Networks, 2004) είναι επίσης μια προσέγγιση που προσπαθεί να παρέχει αποδοτική δρομολόγηση ερωτήσεων πάνω σε σημασιολογικές συνόψεις της γνώσης κόμβων και δεικτοδότηση αυτών των συνόψεων σε έναν κατανεμημένο κατάλογο σε επίπεδο υπέρ-κόμβων με τεχνικές DHT. Τα μεταδεδομένα χρησιμοποιούνται επίσης για σκοπούς ευρετηρίασης. Η δουλειά που παρουσιάζεται στο (Sartiani, Manghi, Ghelli, & Conforti, 2004) περιγράφει ένα μοντέλο υπέρ-κόμβων με XML δεδομένα και μηδενικό κόστος διαχείρισης. Το σχήμα ευρετηρίου που προτείνεται είναι βασισμένο στα μεταδεδομένα (XML Schema) και σε μια ιεραρχική προσέγγιση για τη δημιουργία και διαχείριση ενός καθολικού ευρετηρίου. Τα τελευταία χρόνια υπάρχουν επίσης προσπάθειες που χρησιμοποιούν οντολογίες για τη δημιουργία σημασιολογικών ευρετηρίων σε δίκτυα ομότιμων κόμβων. Τέτοιες προσεγγίσεις παρουσιάζονται στα (Nejdl W. , EDUTELLA: a P2P networking infrastructure based on RDF, May 2002), (Elenius & Ingmarsson, 2004), (Verma), (Castano A. , 2003), (Nejdl W. , 2003) όπου υπάρχουν ευρετήρια βασισμένα σε οντολογίες σαν υποδομή για τη διαχείριση γνώσης στο σημασιολογικό ιστό. Τέλος τα Σημασιολογικά Υπερκείμενα Δίκτυα¹³ που προτείνονται στο (Crespo & Garcia-Molina, Semantic Overlay Networks for P2P Systems, 2004) είναι ένας τρόπος (λογικής)οργάνωσης ενός γνωστικού δικτύου ώστε οι κόμβοι που σχετίζονται σημασιολογικά να δημιουργούν ένα υπερκείμενο

¹² δλδ λίστες με έννοιες ταξινομημένες με βάση κάποιο κριτήριο, συνήθως θεματικό

¹³ Semantic Overlay Networks (SONs)

δίκτυο και οι ερωτήσεις να δρομολογούνται προς το πιο κατάλληλο τέτοιο δίκτυο αυξάνοντας τις πιθανότητες απάντησης.

Κόμβοι οι οποίοι κρατάνε δείκτες προς άλλους κόμβους με παρόμοιο περιεχόμενο, σχηματίζουν ένα σημασιολογικό υπερκείμενο δίκτυο¹⁴. Το Gridvine (Aberer, Cudré-Mauroux, Hauswirth, & Pelt, 2004) παρέχει ένα υπερκείμενο δίκτυο πάνω από το PGrid. Ενώ το PGrid σαν ένα δομημένο ομότιμο δίκτυο για την αποδοτική δρομολόγηση μηνυμάτων παρέχει το «φυσικό» επίπεδο, το Gridvine εισάγει ένα σημασιολογικό επίπεδο πιο πάνω, για τη διαχείριση και αντιστοίχιση δεδομένων και μεταδεδομένων, σαν το «λογικό» επίπεδο. Στην πραγματικότητα, η αποδοτικότητα του αλγορίθμου αναζήτησης οφείλεται όχι στην έξυπνη προώθηση ερωτήσεων στο υπερκείμενο δίκτυο, αλλά με την εφαρμογή της από κάτω DHT προσέγγισης για την αντιστοίχιση των όρων στους κόμβους. Ο στόχος σε αυτό το σημείο είναι ένα αποδοτικός μηχανισμός αναζήτησης βασισμένος στη δρομολόγηση των ερωτήσεων προς τους κόμβους που βρίσκονται πιο κοντά στο περιεχόμενο της ερώτησης. Μια προσέγγιση για να το πετύχει κανείς αυτό είναι μέσω της κατηγοριοποίησης του περιεχομένου κάθε κόμβου σε κάποιο κοινό θεματικό διάνυσμα με κάθε στοιχείο του διανύσματος να περιέχει τη σχετικότητα του κόμβου με το αντίστοιχο θέμα. Ένα παράδειγμα στο οποίο τα έγγραφα του δικτύου είναι οργανωμένα στις δικές τους διανυσματικές αναπαραστάσεις τέτοιες ώστε ο χώρος αναζήτησης για μια ερώτηση να βρίσκεται γύρω από τα σχετικά έγγραφα επιτυγχάνοντας αποδοτικότητα και ακρίβεια, είναι το pSearch (Tang, Xu, & Dwarkadas, 2003). Το χαρακτηριστικό του pSearch είναι η αποδοτικότητα του τρόπου με τον οποίο οι κόμβοι γνωρίζουν τους κοντινούς τους γείτονες. Ένα μειονέκτημά του είναι ότι όλα τα έγγραφα πρέπει να αντιστοιχηθούν στον ίδιο σημασιολογικό χώρο(μικρών διαστάσεων) αναζήτησης και ότι οι διαστάσεις του υπερκείμενου επιπέδου εξαρτώνται αυστηρά από τις διαστάσεις του διανύσματος, με αποτέλεσμα κάθε κόμβος να πρέπει να γνωρίζει πολλούς γείτονες όταν τα διανύσματα έχουν πολλές διαστάσεις. Μια άλλη προσέγγιση (Voulgaris, Kermarrec, Massoulie, & van Steen, 2004) βασίζεται στην ομαδοποίηση με «τυχαίο περίπατο¹⁵» όπου οι κόμβοι με παρόμοιο περιεχόμενο πρέπει να αναγνωριστούν μεταξύ τους. Η υπόθεση είναι ότι οι ερωτήσεις που στέλνονται από τους κόμβους είναι

¹⁴ βλ. προηγούμενη παράγραφο

¹⁵ random walk

σημασιολογικά κοντά με τους ίδιους τους κόμβους. Αυτό έχει σαν αποτέλεσμα να υπάρχει αυξημένη πιθανότητα οι γείτονες ενός κόμβου(οι κόμβοι στην ίδια ομάδα με τον αρχικό κόμβο) να έχουν απάντηση στην ερώτηση. Το πρόβλημα βρίσκεται στο τί πληροφορίες πρέπει να ανταλλάξουν οι κόμβοι μεταξύ τους για να προσδιορίζουν ότι σχετίζονται μεταξύ τους, όταν πρόκειται για αναζητήσεις κειμένου. Όταν δεν υπάρχει κάποια κοινή δομή δεδομένων(όπως ένα συγκεκριμένο σύνολο όρων) την οποία να μπορούν να χρησιμοποιήσουν για να περιγράψουν το περιεχόμενό τους, τότε όλο το περιεχόμενο πρέπει να διαμοιράζεται. Αυτό έχει σαν αποτέλεσμα μεγάλος όγκος δεδομένων να πρέπει να διαμοιράζεται μεταξύ των κόμβων. Στο Freenet (Clarke, Sandberg, Wiley, & Hong, Freenet: A distributed anonymous information storage and retrieval system, 2000) γίνεται αποθήκευση δεικτών προς το δημοφιλές περιεχόμενο με βάση τις προηγούμενες απαντήσεις, όπως είδαμε και στην προηγούμενη παράγραφο. Εν συντομία, όταν ένας κόμβος προωθεί μια ερώτηση σε κάποιον κόμβο ο οποίος επιτυχώς ανακαλεί τα δεδομένα, η διεύθυνσή του περιλαμβάνεται στην απάντηση. Για επιλογή κόμβων βάση εμπειρίας, υπάρχει δουλειά σχετική με «δείκτες δρομολόγησης¹⁶» στους οποίους ο κόμβος διατηρεί γνώση σχετικά με το περιεχόμενο στο οποίο έχει πρόσβαση από τους γείτονές του. Για παράδειγμα στο (Crespo & Garcia-Molina, 2002) οι κόμβοι συνοψίζουν τη γνώση τους σε ένα σύνολο από θέματα και το «διαφημίζουν» μαζί με τον αριθμό των εγγράφων που είναι προσβάσιμα από τους άμεσους γείτονές τους. Λέγοντας «προσβάσιμα» εννοούν είτε ότι ο κόμβος έχει έγγραφα πάνω στο συγκεκριμένο θέμα είτε ότι γνωρίζει άλλους κόμβους που έχουν. Το πρόβλημα με αυτήν τη προσέγγιση είναι πως είτε τα ευρετήρια είναι πολύ μεγάλα(το οποίο έχει μεγάλο κόστος συντήρησης) είτε δεν είναι αρκετά πλούσια ώστε να έχουν επικαλύψεις μεταξύ των κόμβων, με αποτέλεσμα να υπάρχουν αδιέξοδα στη διαδικασία προώθησης. Από την άλλη πλευρά θα θέλαμε να σημειώσουμε πως στην ίδια δουλειά γίνεται εκμετάλλευση της επιλεκτικότητας που εμφανίζουν οι θεματικές περιοχές στη συλλογή των εγγράφων κάθε κόμβου. Πρόκειται για ένα μέγεθος γνωστό ευρέως από τη περιοχή των βάσεων δεδομένων που χρησιμοποιείται και σε αυτήν τη περίπτωση. Παράλληλα ένα ακόμα χαρακτηριστικό είναι η χρήση της κοινής θεματικής ιεραρχίας για την αντιστοίχιση

¹⁶ routing indices

του περιεχομένου κάθε κόμβου. Αυτό αποτελεί μάλιστα βασικό χαρακτηριστικό αρκετών ακόμα εργασιών. Από τη μια πλευρά η ανάγκη της περιγραφής είναι προφανής από την άλλη όμως μια απλή θεματική ιεραρχία πρώτον έχει περιορισμένη σημασιολογική δύναμη και δεύτερον περιορίζεται τις περισσότερες φορές στην απλή εκμετάλλευση(λεκτικά) των όρων που εμφανίζονται στα εκάστοτε έγγραφα ή συλλογές. Η περίπτωση των σημασιολογικών υπερκείμενων δικτύων που είχαμε δει και νωρίτερα επιτρέπουν στους κόμβους να περιγράψουν το περιεχόμενό τους βάσει ενός κοινού συνόλου από όρους. Τις περισσότερες φορές αυτοί οι όροι είναι οργανωμένοι σε ένα δίκτυο ή μια ιεραρχία θεμάτων επιτρέποντας τον προσδιορισμό της σημασιολογικής ομοιότητας μεταξύ τους. Ένα σύνολο από θεματικές περιοχές χαρακτηρίζουν τον τομέα εμπειρίας κάθε κόμβου. Επίσης, κάθε κόμβος γνωρίζει για τους υπόλοιπους αναλύοντας τα «διαφημιστικά» μηνύματα (Haase, Siebes, & van Harmelen, Peer selection in peer-to-peer networks with semantic topologies, 2004) και τις απαντήσεις (Tempich, Staab, & Wranik, 2004). Με αυτόν τον τρόπο οι κόμβοι δημιουργούν ομάδες με βάση τις σχετιζόμενες σημασιολογικές περιγραφές. Δοθείσας μιας ερώτησης μια μετρική για την απόσταση επιτρέπει την προώθηση των ερωτήσεων¹⁷ σε κόμβους που σχετίζονται σημασιολογικά με αυτές. Η εκμετάλλευση της σημασιολογίας είναι και σε αυτή την περίπτωση σημαντική τόσο στο κομμάτι της περιγραφής των δεδομένων όσο και στο κομμάτι της σύγκρισης ή αντιστοίχισης των περιγραφών (δεδομένων και κόμβων) για τον προσδιορισμό της μεταξύ τους απόστασης. Ιδιαίτερα χρήσιμη προς αυτή την κατεύθυνση είναι η χρήση των οντολογιών οι οποίες εξ ορισμού αποτελούν ένα ισχυρό εργαλείο αποτύπωσης της σημασιολογίας σε οποιαδήποτε περιοχή.

Συστήματα για αναζήτηση με τη χρήση οντολογιών σαν το KAON (Maedche, Motik, Stoljanovic, Studer, & Volz, 2003) βασίζονται σε οργανωμένες συλλογές οντολογιών¹⁸. Στην ίδια κατηγορία και αντιμετωπίζοντας το πρόβλημα της αντιστοίχισης σχημάτων ανήκουν τα (Castano, Ferrara, & Montanelli, H-match: an Algorithm for Dynamically Matching Ontologies in Peer-based Systems, 2003), (Castano, Ferrara, Montanelli, Pagani, & Rossi, 2003). Μερικά συστήματα

¹⁷ Είναι σημαντικό ότι και οι ερωτήσεις θα πρέπει να κάνουν χρήση για την περιγραφή τους του κοινού συνόλου των όρων που υπάρχουν στο σύστημα.

¹⁸ Αναφερόμαστε σε «αποθήκες» οντολογιών – Ontology Registries

ασχολούνται με το πρόβλημα της δρομολόγησης ερωτήσεων (Crespo & Garcia-Molina, Routing Indices for Peer-to-Peer Systems, 2002) (Crespo & Garcia-Molina, Semantic Overlay Networks for P2P Systems, 2004) αλλά είτε δεν έχουν την ευελιξία να χειριστούν σύνθετες ερωτήσεις πάνω στο περιεχόμενο ετερογενών μοντέλων είτε αδυνατούν να καλύψουν τις ειδικές απαιτήσεις που επιβάλλει η χρήση οντολογιών. Παρ' όλα αυτά βασίζονται σε ενδιαφέρουσες προσεγγίσεις που θα μπορούσαν να επεκταθούν και για την περίπτωση μας. Αυτές οι προσεγγίσεις βασίζονται στο υπερκείμενο σημασιολογικό δίκτυο και προς την ίδια κατεύθυνση υπάρχει αρκετή βιβλιογραφία σήμερα. Άλλα παραδείγματα συστημάτων είναι τα (Li, Lee, Sivasubramaniam, & Lee, 2004) και (Löser, Staab, & Tempich, Semantic Methods for P2P Query Routing, 2005). Το πρώτο ασχολείται με τεχνικές σημασιολογικής ομαδοποίησης για τη δημιουργία ενός σημασιολογικού «μικρόκοσμου¹⁹», ενώ το δεύτερο είναι μια προσέγγιση για προώθηση ερωτήσεων που εκμεταλλεύεται τη σημασιολογική ομοιότητα μεταξύ των ερωτήσεων και θεμάτων για τη δημιουργία ενός σημασιολογικού ευρετηρίου. Τέλος, υπάρχουν και κάποιες μετρικές για την αξιολόγηση τέτοιων δικτύων. Και στις δύο δουλειές υπάρχουν και πειράματα μαζί με τα αποτελέσματά τους ενώ κάποιες επιπλέον μετρικές παρουσιάζονται στο (Schmitz & Loser, 2006).

1.5.2. Ευρετήρια σε δίκτυα Ομότιμων Κόμβων και Σημασιολογικά Υπερκείμενα Δίκτυα

Σε συνέχεια της προηγούμενης παραγράφου θα παραθέσουμε κάποιες πιο πρόσφατες δουλειές που αναφέρονται σε επίσης δίκτυα ομότιμων κόμβων με λίγο περισσότερη έμφαση είτε στη δημιουργία ευρετηρίων σε τέτοια περιβάλλοντα είτε στην εκμετάλλευση ενός υπερκείμενου δικτύου.

Δύο σχετικές εργασίες είναι οι (Crainiceanu, Linga, Gehrke, & Shanmugasundaram, 2004) και (Tang & Dworkadas, 2004). Στην πρώτη περίπτωση έχουμε την πρόταση για ένα αποδοτικό ευρετήριο, το P-tree, για XML δεδομένα με στόχο την αποδοτική επεξεργασία και προώθηση ερωτήσεων. Πρόκειται για μια σύνθετη δομή και έναν αλγόριθμο που υποστηρίζουν αναζητήσεις (τόσο ερωτήσεων εύρους όσο και ισότητας), εισαγωγές και διαγραφές στο δίκτυο με κατά μέσο όρο λογαριθμικό

¹⁹ Είναι τα γνωστά συστήματα “small world”

αριθμό μηνυμάτων σε σχέση με τον αριθμό των κόμβων. Στη δεύτερη περίπτωση υπάρχει η πρόταση για μια ολόκληρη αρχιτεκτονική ομότιμων κόμβων και ένα σύστημα, το eSearch, το οποίο αναλαμβάνει τη δημιουργία ευρετηρίου πάνω στα έγγραφα του δικτύου βασιζόμενο όμως καθαρά πάνω στους όρους που περιέχονται στα έγγραφα. Προς την ίδια κατεύθυνση με τις προηγούμενες είναι η δουλειά των (Sartiani, Marciniak, & Cerroni, XPeer: A Self-organizing XML P2P Database System, 2004) όπου το XPeer είναι μια προτεινόμενη αρχιτεκτονική ενός συστήματος διαχείρισης XML δεδομένων σε περιβάλλον ομότιμου δικτύου. Το σύστημα βασίζεται σε μια κοινή βάση δεδομένων ενώ σαν γλώσσα ερωτήσεων χρησιμοποιείται η XQuery. Επίσης η αρχιτεκτονική του συστήματος βασίζεται σε ένα δίκτυο από υπέρ-κόμβους. Μια άλλη εργασία σχετική πάλι με τη δημιουργία ευρετηρίων είναι η (Rostami, Habibi, Abolhassani, Amirkhani, & Rahnama, 2006), όπου γίνεται εκμετάλλευση της σημασιολογίας με τη χρήση οντολογιών, χωρίς όμως να γίνεται σαφής προσδιορισμός της γλώσσας περιγραφής ή των ερωτήσεων ενώ δεν υπάρχει και συζήτηση για το θέμα της ετερογένειας μεταξύ των κόμβων. Τέλος η εφαρμογή γίνεται σε συγκεκριμένη τοπολογία όπως αυτή του Hypercup (Schlosser M. , Sintek, Decker, & Nejd, 2002). Στην ίδια δουλειά υπάρχει αναφορά και στη δρομολόγηση ερωτήσεων με την οποία όμως θα ασχοληθούμε εκτενέστερα στην επόμενη παράγραφο. Στην ίδια κατηγορία θα συμπεριλάβουμε τρεις ακόμη εργασίες, αυτές των (Li & Vuong, 2005) , (Sangpachatanaruk & Znati, 2004) και (Abidi & Pang, 2004). Στην πρώτη έχουμε την περίπτωση μιας DHT προσέγγισης με χρήση όμως οντολογιών. Τα χαρακτηριστικά των οντολογιών χρησιμοποιούνται για ομαδοποίηση των κόμβων ενώ οι ερωτήσεις προωθούνται με χρήση του DHT. Στη δεύτερη εργασία έχουμε και πάλι μια DHT προσέγγιση με χρήση όμως αυτή τη φορά μιας κοινής οντολογίας ενώ στη τρίτη, υπάρχει μια οντολογία για την περιγραφή των κόμβων και τη δημιουργία ενός υπερκείμενου δικτύου με βάση αυτές τις περιγραφές. Σε αυτήν την εργασία ενώ υπάρχει δουλειά για την αντιμετώπιση της ετερογένειας μεταξύ των κόμβων δεν υπάρχει ένα ολοκληρωμένο πλαίσιο που να δίνει τη δυνατότητα ισχυρών ερωτήσεων και πλούσιας διαμοιρασμένης σημασιολογικής περιγραφής. Θα κλείσουμε αυτή την παράγραφο με δύο δημοσιεύσεις που αναφέρονται σε ένα σύνολο από εργασίες στην περιοχή στην οποία αναφερόμαστε. Η πρώτη (Joseph & Hoshiai, Decentralized Meta-Data Strategies: Effective Peer-to-Peer Search, 2003) ασχολείται με μια έρευνα στα

συστήματα που υπάρχουν για διαχείριση των μεταδεδομένων και προσπαθεί να εντοπίσει κοινά στοιχεία και διαφορές μεταξύ τους. Στη δεύτερη παραθέτουμε τη λίστα με όλα τα πεπραγμένα του συνεδρίου για την Μετατροπή της Γνώσης στα πλαίσια του Σημασιολογικού Ιστού (Proceedings, 2002). Σε αυτά υπάρχουν ενδιαφέρουσες εργασίες τόσο για τη μοντελοποίηση και σημασιολογική περιγραφή της γνώσης όσο και για το κομμάτι της ετερογένειας μεταξύ διαφορετικών πηγών, όπου περιλαμβάνεται και η αντιστοίχιση μεταξύ οντολογιών.

1.5.3. Δρομολόγηση Ερωτήσεων

Με την παράγραφο αυτή θα ολοκληρώσουμε την παρουσίαση των σχετικών εργασιών που σχετιζόντουσαν άμεσα με τη δημιουργία ενός δικτύου και τα διάφορα προβλήματα γύρω από αυτό. Ένα από τα γνωστά προβλήματα είναι αυτό της δρομολόγησης ερωτήσεων ή αλλιώς, της στρατηγικής για την καλύτερη επιλογή κόμβων για την προώθηση των ερωτήσεων. Σε αυτή τη κατηγορία ανήκουν χαρακτηριστικές εργασίες σαν αυτές που είδαμε στην παράγραφο 1.5.1 (Crespo & Garcia-Molina, Routing Indices for Peer-to-Peer Systems, 2002) ή το σύστημα pSearch που είδαμε στην προηγούμενη παράγραφο. Μια ενδιαφέρουσα εργασία είναι αυτή των (Brunkhorst, Dhraief, Kemper, Nejd, & Wiesner, 2003), όπου προτείνουν κάποια συγκεκριμένη τοπολογία και αρχιτεκτονική υπέρ-κόμβων. Συγκεκριμένα, χρησιμοποιούν κατανεμημένους πίνακες δρομολόγησης και επεξεργασία των ερωτήσεων μέσω τελεστών που τοποθετούν δίπλα στις πηγές των δεδομένων. Προς την ίδια κατεύθυνση κινούνται τόσο το σύστημα Neurogrid (Joseph, NeuroGrid: Semantically Routing Queries in Peer-to-Peer Networks, 2002) όσο και η δουλειά των (Hose, Karnstedt, Sattler, & Stehr, 2005). Στο Neurogrid γίνεται σημασιολογική δρομολόγηση των ερωτήσεων βάση των περιγραφών που έχει κάθε χρήστης για τους κόμβους. Οι περιγραφές αυτές προσαρμόζονται βάσει των μεταδεδομένων που διακινούνται στο δίκτυο. Για το σκοπό αυτό υπάρχει ένα μηχανισμός εκμάθησης που αναλαμβάνει την ενημέρωση των αντίστοιχων περιγραφών. Στη δεύτερη εργασία υπάρχει ένα σύστημα διαχείρισης των δεδομένων των κόμβων. Συγκεκριμένα, χτίζονται «φίλτρα δρομολόγησης» με βάση τα αποτελέσματα των ερωτήσεων που διακινούνται στο δίκτυο. Μια ενδιαφέρουσα και ολοκληρωμένη εργασία είναι αυτή των (Haase, Siebes, & van Harmelen, 2007). Σκοπός είναι η επιλογή κόμβων για την προώθηση των ερωτήσεων βάσει εμπειρίας.

Χρησιμοποιείται συγκεκριμένη στρατηγική για τη διαφήμιση της γνώσης των κόμβων ενώ γίνεται εφαρμογή σε κάποιο συγκεκριμένο σενάριο τήρησης βιβλιογραφίας. Στο σύστημα SIMPEER (Doulkeridis, Vlachou, Kotidis, & Vazirgiannis, 2007) παρουσιάζεται ένα ολόκληρο πλαίσιο για αναζήτηση σε ομότιμα δίκτυα, βάσει ομοιότητας. Το πλαίσιο περιλαμβάνει ομαδοποίηση των κόμβων χρησιμοποιώντας κατανεμημένα στατιστικά με τη μορφή ιστογραμμάτων. Αυτά διατηρούνται από τους υπέρ-κόμβους και βοηθούνε τον εντοπισμό των κοντινότερων κόμβων. Τέλος θα πρέπει προσθέσουμε σε αυτήν τη παράγραφο ανήκουν και εργασίες που ασχολούνται με δρομολόγηση πάνω από υπερκείμενα δίκτυα (Löser, Staab, & Tempich, Semantic Social Overlay Networks, 2007) στις οποίες, όπως είδαμε και στην παράγραφο 1.5.1, υπάρχει μια στροφή πιο πρόσφατα και οι οποίες είναι σημαντικές γιατί εκμεταλλεύονται το σημασιολογικό κομμάτι, που μας ενδιαφέρει.

1.5.4. Ευρετήρια και συνόψεις πάνω σε XML δεδομένα

Με τις προηγούμενες παραγράφους του κεφαλαίου 1.5 καλύψαμε τα θέματα που έχουν να κάνουν με τη διαχείριση της γνώσης σε επίπεδο δικτύου, αντιμετωπίζοντας τόσο το θέμα της ίδιας της δημιουργίας του δικτύου, όσο και της επεξεργασίας και δρομολόγησης των ερωτήσεων στο ίδιο επίπεδο. Στη δικιά μας περίπτωση όπως είδαμε και από την εισαγωγή, επικεντρώσαμε σε δίκτυα που είχαν δεδομένα μορφής XML και μάλιστα μας απασχόλησε περισσότερο το θέμα τις επεξεργασίας των ερωτήσεων. Στην τρέχουσα παράγραφο θα δούμε βιβλιογραφία σχετική με την επεξεργασία και εκμετάλλευση XML δεδομένων ενώ στο υπόλοιπο του κεφαλαίου, έχοντας ολοκληρώσει με το υπόβαθρο που αποκτήσαμε, θα δούμε τις πιο σχετικές εργασίες με το αντικείμενο της εργασίας μας.

Στην περιοχή αυτή, οι πιο ενδιαφέρουσες δουλειές που εντοπίσαμε ήταν αυτές των Α. Πολυζώτη και Μ. Γαροφαλάκη (Polyzotis & Garofalakis, Structure and Value Synopses for XML Data Graphs, 2002), (Polizotis & Garofalakis, 2006), (Polyzotis & Garofalakis, Statistical Synopses for GraphStructured XML Databases, 2002). Αυτές κινούνται στην περιοχή των συνόψεων XML δεδομένων προτείνοντας συγκεκριμένες δομές(XSketches) και μεθοδολογίες για τη σύνοψη από τη μια

πλευρά των δεδομένων και την επεξεργασία από την άλλη εκφράσεων διαδρομών²⁰ με αποδοτικό τρόπο. Πρακτικά οι συνόψεις ικανοποιούν την ανάγκη της εκμετάλλευσης των στατιστικών στοιχείων για ολόκληρη τη συλλογή των εγγράφων αλλά με πολύ αποδοτικό τρόπο. Μιλώντας για στατιστικά από δω και πέρα, θα αναφερόμαστε κατά κύριο λόγο στο μέγεθος της επιλεκτικότητας των δεδομένων. Οι συνόψεις είναι σε μορφή γράφου και προκύπτουν από την επεξεργασία των αρχικών δεδομένων ενώ προτείνεται και μεθοδολογία για την επεξεργασία των ερωτήσεων με τη μορφή εκφράσεων διαδρομών απλών ή σύνθετων. Ανάλογη δουλειά, είναι αυτή των (Chan, Felber, Garofalakis, & Rastogi, 2002) όπου έχουμε κάποια δομή που λειτουργεί σαν ευρετήριο για XML δεδομένα και είναι κατασκευασμένη ώστε να υποστηρίζει φιλτράρισμα των εγγράφων με βάση πάλι εκφράσεις διαδρομών. Μια ακόμη εργασία για τη δημιουργία ενός ευρετηρίου ερωτήσεων είναι των (Diao, Altinel, Franklin, Zhang, & Fischer, 2003). Σε αυτήν προτείνεται το σύστημα XFilter το οποίο στηρίζεται σε μια μηχανή πεπερασμένων καταστάσεων για το φιλτράρισμα XML εγγράφων. Επιπλέον υπάρχει μια προσέγγιση για το συνδυασμό πολλών εκφράσεων διαδρομών σε μια αυτόματη μηχανή καταστάσεων.

1.5.5. Εκτίμηση Επιλεκτικότητας

Ήδη από τις προηγούμενες παραγράφους είδαμε τη σημαντικότητα της επιλεκτικότητας ως μεγέθους για την σύγκριση των διαφόρων πηγών δεδομένων μεταξύ τους αλλά και ως κριτηρίου για τη δρομολόγηση των ερωτήσεων σε ένα δίκτυο. Επίσης όπως υπαινιχτήκαμε και στην παράγραφο 1.5.1 πρόκειται για το γνωστό μέγεθος από την περιοχή των βάσεων δεδομένων όπου βέβαια η χρήση του είναι γνωστή εδώ και πολλά χρόνια. Το καινούριο στοιχείο είναι η ανάγκη υπολογισμού της επιλεκτικότητας για XML δεδομένα ή ακόμα πιο εξειδικευμένα για σημασιολογικά δεδομένα που βρίσκονται περιγεγραμμένα με τη βοήθεια οντολογιών και κάποιας συγκεκριμένη γλώσσας όπως η OWL. Μια οντολογία, όπως είδαμε και στην εισαγωγή, αποτελεί ένα κοινά αποδεκτό εννοιολογικό προσδιορισμό κάποιας περιοχής. Για να είναι αυτό άμεσα εκμεταλλεύσιμο θα πρέπει να είναι περιγεγραμμένο και με κάποιον κοινά αποδεκτό τρόπο. Ένας

²⁰ path expressions

τέτοιος τρόπος σήμερα, είναι το πρότυπο της OWL. Αυτά αποτέλεσαν μια από τις ερευνητικές προκλήσεις της εργασίας αυτής και πρόκειται και για ένα από τα σημεία συνεισφοράς της. Στην παράγραφο αυτή θα ανατρέξουμε στις εργασίες που είναι σχετικές με την εκτίμηση της επιλεκτικότητας είτε από την παλιότερη βιβλιογραφία και την περιοχή των σχεσιακών βάσεων δεδομένων είτε πιο πρόσφατες προσεγγίσεις που ασχολούνται και με XML δεδομένα. Φυσικά στην ίδια κατηγορία θα μπορούσαμε να βάλουμε και εργασίες που είδαμε στην παράγραφο 1.5.4 αφού η σύνοψη των δεδομένων αποτελεί πολλές φορές το πρώτο βήμα για την αποθήκευση στατιστικών που θα χρησιμοποιηθούν στην συνέχεια για τον υπολογισμό της επιλεκτικότητας. Επίσης θα θέλαμε να δώσουμε έμφαση στο γεγονός ότι ένα κοινό στοιχείο που συναντά κανείς στις σχετικές εργασίες είναι η απόφαση για υιοθέτηση ή όχι της υπόθεσης της ανεξαρτησίας των δεδομένων. Μιλώντας για την ανεξαρτησία των δεδομένων αναφερόμαστε στην υπόθεση για λόγους απλούστευσης ότι η εμφάνιση διαφορετικών στοιχείων δεδομένων μέσα σε μια συλλογή, είναι ανεξάρτητη μεταξύ τους. Περισσότερα για όλα αυτά θα δούμε και στο 4 κεφάλαιο[§4.2.1]. Τέλος, εργασίες που αναφέρονται πιο συγκεκριμένα στην εκτίμηση του κόστους μιας ερώτησης, με χρήση της επιλεκτικότητας, θα τις δούμε στην παράγραφο 1.5.6.

Θα ξεκινήσουμε με τρεις εργασίες που αναφέρονται σε σχεσιακές βάσεις δεδομένων. Στην πρώτη, οι (Poosala & Ioannidis, 1997) προσπαθούν να αποφύγουν την υπόθεση της ανεξαρτησίας των δεδομένων προτείνοντας δυο εναλλακτικές: στην πρώτη περίπτωση κάνουν χρήση πολυδιάστατων ιστογραμμάτων ενώ στη δεύτερη περίπτωση χρησιμοποιούν τεχνικές από τη γραμμική άλγεβρα όπως η SVD²¹ πάνω στα χαρακτηριστικά των σχέσεων. Η επόμενη εργασία (Bruno & Chaudhuri, 2004) ασχολείται με τον υπολογισμό της επιλεκτικότητας για συγκεκριμένες ερωτήσεις. Η δουλειά βασίζεται σε έναν συγκεκριμένο αλγόριθμο δυναμικού προγραμματισμού ενώ γύρω από αυτόν αναπτύχθηκε ένα ολόκληρο πλαίσιο το οποίο κάνει χρήση των πλάνων των ερωτήσεων και υπολογισμού της επιλεκτικότητας υπό συνθήκες. Η τρίτη εργασία αναφέρεται ουσιαστικά με εκμετάλλευση της επιλεκτικότητας για επανεγγραφή των ερωτήσεων με στόχο τη βελτιστοποίηση της εκτέλεσής τους. Η μεθοδολογία ασχολείται με ερωτήσεις

²¹ Singular Value Decomposition

ένωσης ενώ περιλαμβάνεται φυσικά και υπολογισμός του κόστους των ερωτήσεων. Ξεχωριστή αναφορά θα πρέπει να κάνουμε στην εργασία του Σ. Χριστοδουλάκη (Christodoulakis, Implications of Certain Assumptions in Database Performance Evaluation, 1984), (Christodoulakis, Estimating Record Selectivities, 1983) ο οποίος έχει ασχοληθεί εκτεταμένα με το θέμα της υπόθεσης της ανεξαρτησίας στην εκτίμηση της απόδοσης των βάσεων δεδομένων και στη συγκεκριμένη, περιλαμβάνεται μια από τις πιο στιβαρές αποδείξεις για την απόρριψή της. Φυσικά η απόδειξη αναφέρεται πάλι σε σχεσιακά δεδομένα όμως η αδυναμία της υπόθεσης φαίνεται να υπάρχει ακόμα πιο πολύ σε XML δεδομένα χωρίς να υπάρχει κάποια επίσημη, σύγχρονη απόδειξη.

Περνώντας στις εργασίες που ασχολούνται με XML δεδομένα, υπάρχουν κάποιες που σχετίζονται πιο άμεσα με όσα είδαμε στην παράγραφο 1.5.4 όπως η (Polyzotis, Garofalakis, & Ioannidis, Selectivity Estimation for XML Twigs, 2004). Εκεί εκμεταλλεύονται το γνωστό μοντέλο του XSketch για την απόκτηση ακριβών συνόψεων πάνω στα δεδομένα. Στη συνέχεια υπάρχει μια αναπαράσταση των ενώσεων(ερωτήσεων) στον πολυδιάστατο χώρο και μια μεθοδολογία για τον υπολογισμό της επιλεκτικότητας σύνθετων ερωτήσεων διαδρομών με βάση τα στατιστικά των περιλήψεων. Μια προσέγγιση προς την ίδια κατεύθυνση είναι το γνωστό σύστημα XPathlearner (Lim, Wang, Padmanabhan, Vitter, & Parr, 2002). Το σύστημα συγκεντρώνει στατιστικά στοιχεία βάσει των απαντήσεων των ερωτήσεων κατασκευάζοντας ένα Markov ιστόγραμμα των τιμών επιλεκτικότητας των διαφόρων διαδρομών. Μια ανάλογη δουλειά είναι αυτή των (Wang, Wang, Meng, & Wang, 2004). Σε αυτήν την περίπτωση γίνεται χρήση δισδιάστατων ιστογραμμάτων πάνω στα οποία ορίζεται ένα σύνολο από λειτουργίες. Στη συνέχεια δημιουργείται ένα δέντρο με τα κόστη των διαφόρων κόμβων ενώ ο υπολογισμός του κόστους μιας οποιασδήποτε έκφρασης διαδρομής υπολογίζεται με εκτέλεση του αντίστοιχου δέντρου κόστους. Το (Soldak, 2005) ασχολείται περισσότερο με την αποθήκευση των τιμών επιλεκτικότητας σε κάποιον γράφο ενώ στο (Aboulnaga, Alameldeen, & Naughton, 2001) ασχολούνται με τη σύνοψη των δεδομένων προτείνοντας δύο τεχνικές μια με χρήση δέντρων μονοπατιών και μία με χρήση Markov πινάκων. Μια πρόσφατη και εκτεταμένη παρουσίαση για την εκτίμηση του μεγέθους δομικών ενώσεων πάνω σε XML βάσεις δεδομένων υπάρχει στο (Luo, Jiang, Hou, Yan, & Zhu, 2008). Οι ενώσεις μοντελοποιούνται σαν ενώσεις

ισότητας με σχεσιακό τρόπο και υπάρχει εκτεταμένη θεωρητική ανάλυση ώστε γνωστές τεχνικές για τη σύνοψη δεδομένων, να εφαρμοστούν και σε αυτήν την περίπτωση. Τεχνικές όπως κυματικοί μετασχηματισμοί και sketches. Τέλος μια ενδιαφέρουσα εργασία (Getoor, Taskar, & Koller, 2001) ασχολείται με την εκτίμηση της επιλεκτικότητας με χρήση πιθανοτικών μοντέλων. Ουσιαστικά υπάρχει αναπαράσταση των δεδομένων σε ένα Bayesian δίκτυο και εκμετάλλευση της αντίστοιχης θεωρίας πιθανοτήτων για εκτίμηση της επιλεκτικότητας υπό συνθήκη.

1.5.6. Εκτίμηση Κόστους

Οι εργασίες που θα δούμε σε αυτήν την παράγραφο σχετίζονται άμεσα με όσες είδαμε στην προηγούμενη(1.5.5) απλά οι περισσότερες εστιάζονται περισσότερο στο θέμα της εκτίμησης του κόστους των ερωτήσεων. Σε αυτές μπορεί κανείς να βρει πιο παλιές προσεγγίσεις για σχεσιακά συστήματα βάσεων δεδομένων όπως οι (Piatetsky-Shapiro & Connell, 1984), (Selinger, Astrahan, Chamberlin, Lorie, & Price, 1979). Και οι δύο εργασίες έχουν σαν στόχο τη βελτιστοποίηση των ερωτήσεων και ήταν χρήσιμες στην κατανόηση της σημασίας της έννοιας της επιλεκτικότητας με στόχο την εκτίμηση του κόστους διαφόρων τύπων ερωτήσεων. Σε μια παλιότερη έρευνα (Graefe, 1993) υπάρχει μια εκτεταμένη αναφορά στο θέμα της επεξεργασίας των ερωτήσεων και της εκτίμησης του κόστους τους και ο αναγνώστης έχει την ευκαιρία να ανατρέξει σε όλες τις διαφορετικές μεθοδολογίες που υπάρχουν στην περιοχή. Μια εργασία που έρχεται πιο κοντά στο πλαίσιο των οντολογιών που μας ενδιαφέρει είναι αυτή των (Ruckhaus, Ruiz, & Vidal, 2008). Αυτοί προτείνουν ένα μοντέλο κόστους και μια τεχνική δειγματοληψίας για την εκτίμηση των δεδομένων που δεν είναι γνωστά εκ των προτέρων και σχετίζονται με την επιλεκτικότητα πάνω σε μια συγκεκριμένη δομή αναπαράστασης της οντολογίας όπως την περιγράφουν οι ίδιοι. Δύο εργασίες από την ίδια ομάδα { (Gravano, Garcia-Molina, & Tomasic, Precision and Recall of GLOSS Estimators for Database Discovery, 1994?), (Gravano, Garcia-Molina, & Tomasic, The Effectiveness of Gloss for the Text Database Discovery Problem, 1994?) }, ασχολούνται με την εκτίμηση του κόστους πάνω σε βάση δεδομένων συλλογής κειμένων με πολύ χρήσιμες όμως προσεγγίσεις πάνω στους τύπους και τη μεθοδολογία εκτίμησης του κόστους κάνοντας νύξεις και για το θέμα της ανεξαρτησίας των δεδομένων. Τέλος η πιο σχετική δουλειά με την εργασία μας είναι αυτή των (Stocker, Seaborne, & Bernstein,

2008). Αυτοί ασχολούνται με το θέμα της εκτίμησης της επιλεκτικότητας με στόχο τη βελτιστοποίηση SPARQL εκφράσεων. Ο αλγόριθμος αναπτύσσεται πάνω σε γράφο προτύπων και προτείνονται μια σειρά από ευρηστικοί κανόνες για τον υπολογισμό σύνθετων επιλεκτικότητας σε περιπτώσεις ενώσεων. Η δουλειά βέβαια δεν είναι ολοκληρωμένη αλλά κινείται προς την κατεύθυνση που μας ενδιαφέρει για την εκμετάλλευση στοιχείων της οντολογίας με σκοπό των υπολογισμό σύνθετων επιλεκτικότητας.

1.5.7. Υπολογισμός Επιλεκτικότητας σε Αντικειμενοστρεφείς Βάσεις Δεδομένων

Μια σειρά από εργασίες που προέρχονται από διαφορετική περιοχή αλλά έχουν να αντιμετωπίσουν τα ίδια προβλήματα σε διαφορετικό απλά περιβάλλον, είναι αυτές από την περιοχή των αντικειμενοστρεφών βάσεων δεδομένων. Το κοινό στοιχείο τους είναι η χρήση και επεξεργασία εκφράσεων διαδρομών. Οι (Gardarin, Gruser, & Tang, 1996) ασχολούνται με αυτό το θέμα προτείνοντας ένα μοντέλο κόστους διάσχισης των μονοπατιών ενώ οι (Stocker, Seaborne, & Bernstein, 2008) προτείνουν μια κατάλληλη γλώσσα ερωτήσεων και την υλοποίηση τεχνικών για τη βελτιστοποίηση των ερωτήσεων. Η υλοποίηση περιλαμβάνει μια μηχανή εκτέλεσης των ερωτήσεων και κάνει χρήση της SPARQL. Το (Bertino, 1994) ασχολείται με μηχανισμούς πρόσβασης για τον υπολογισμό εμφωλευμένων χαρακτηριστικών μέσα σε ένα μονοπάτι. Συγκεκριμένα προτείνουν τον βέλτιστο τρόπο για τη δημιουργία ευρετηρίου για ένα μονοπάτι μην έχοντας βέβαια προτάσεις για περιπτώσεις σύνθετων μονοπατιών. Η τελευταία εργασία (Cho, Hong, & Loh, 2007) έχει σαν στόχο επίσης τον υπολογισμό της επιλεκτικότητας σε εμφωλευμένα χαρακτηριστικά. Το διαφορετικό στοιχείο σε αυτήν την προσέγγιση είναι η εκμετάλλευση των σχέσεων (πολλαπλότητας και συμμετοχής) μεταξύ των αντικειμένων.

1.5.8. Σημασιολογική Ομοιότητα και Αντιστοίχιση Οντολογιών

Έχοντας σαν υπόβαθρο τα προηγούμενα και δεδομένου ότι μέσα στους στόχους μας ήταν η εκμετάλλευση των σημασιολογικών περιγραφών με χρήση των οντολογιών, ανατρέξαμε σε μια σειρά από εργασίες που αντικείμενο είχαν τον

εντοπισμό σημασιολογικών ομοιοτήτων και συσχετίσεων. Στην ίδια κατηγορία ανήκαν και εργασίες που αναφέρονταν σε αντιστοίχιση μεταξύ οντολογιών.

Η πρώτη από αυτές (Castano, De Antonellis, & De Capitani, 2001) έχει σαν στόχο την αντιμετώπιση της ετερογένειας μεταξύ των πηγών δεδομένων. Γίνεται παρουσίαση μιας μεθοδολογίας ενοποίησης για την αναγνώριση στοιχείων από ετερογενή σχήματα. Η μέθοδος βασίζεται σε ένα κοινό σχήμα αναπαράστασης των δεδομένων και τεχνικές για τη μέτρηση της σημασιολογικής απόστασης των αντικειμένων μεταξύ τους. Μια έρευνα για τις διάφορες τεχνικές που είχαν αναπτυχθεί μέχρι εκείνη τη στιγμή για την αντιστοίχιση μεταξύ σχημάτων και ταξινόμησή τους, υπάρχει στο (Rahm & Bernstein, A survey of approaches to automatic schema matching, 2001). Όσον αφορά την αντιστοίχιση μεταξύ οντολογιών, μια κατηγορία εργασιών ασχολούνται κατά βάση με τη δομική ομοιότητα που παρουσιάζουν τα σχήματα μεταξύ τους. Τέτοιες είναι είτε γενικού ενδιαφέροντος εργασίες όπως η (Schickel & Faltings, 2007) είτε επικεντρωμένες σε οντολογίες συγκεκριμένων θεματικών περιοχών όπως στο (Al-Mubaid & Nguyen, 2007). Σε άλλες περιπτώσεις (Thanh Le, Dieng-Kuntz, & Gandon, 2004) γίνεται χρήση συνδυασμένης μετρικής για τον υπολογισμό της ομοιότητας μεταξύ των εννοιών λαμβάνοντας υπόψη τόσο δομικά όσο και λεκτικά χαρακτηριστικά τους.

Μια σημαντική προσέγγιση που θα μας απασχολήσει στη συνέχεια της παραγράφου αλλά αποτέλεσε και οδηγό για το αντίστοιχο θεωρητικό κομμάτι της εργασίας [§Error! Reference source not found.], είναι αυτή που προέκυψε από την επιστημονική περιοχή της Θεωρίας Πληροφοριών²². Όλη η περιοχή στηρίζεται στις βάσεις που έβαλε πρώτος ο Shannon (Shannon, 1948) από πολύ νωρίς. Την ανάλυση της θεωρίας θα την δούμε στην παράγραφο [§4.3.1]. Εργασίες που έχουν στηριχθεί σε αυτήν είναι πρώτα απ' όλα του Resnik στις (Resnik, Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, 1998), (Resnik, Using Information Content to Evaluate Semantic Similarity in a Taxonomy, 1995). Σε αυτές γίνεται αποτελεσματική χρήση της έννοιας του πληροφοριακού περιεχομένου η οποία ορίζεται για την περίπτωση των ιεραρχιών εννοιών και επεκτείνεται μαθηματικά για να καλύψει τον υπολογισμό του πληροφοριακού περιεχομένου συνδυασμού

²² Information Theory

εννοιών που ανήκουν στην ίδια ιεραρχία. Προτείνονται διάφορες παραλλαγές του αρχικού τύπου και οι εφαρμογές περιλαμβάνουν κυρίως ιεραρχίες ή ταξινομίες λέξεων. Για την αξιολόγηση μάλιστα γίνεται χρήση παλιότερων αποτελεσμάτων από εφαρμογές πάνω στο Wordnet. Υπάρχει ο υπαινιγμός για την επέκταση στη περίπτωση διαφορετικής βαρύτητας των εννοιών ανάλογα με το βαθμό συσχέτισης, δεν βρήκαμε όμως κάποια συγκεκριμένη πρόταση ενώ όλες οι εργασίες που θα δούμε και στη συνέχεια αναφέρονται αποκλειστικά σε ιεραρχίες ή σε οντολογίες που διαθέτουν μόνο τη σχέση της κληρονομικότητας²³. Πολύ ενδιαφέρουσες προεκτάσεις είναι αυτές των (Lin, 1998) και (Formica, 2007). Στην πρώτη μπορεί κανείς να βρει μια επέκταση του τύπου του Resnik και τις παραλλαγές της μαζί με εφαρμογές σε διαφορετικούς τύπους προβλημάτων ενώ στη δεύτερη υπάρχει πρόταση για να συμπεριληφθούν στον υπολογισμό τόσο δομικά χαρακτηριστικά όσο και οι τύποι των εννοιών. Υπάρχουν πάλι ενδιαφέρουσες εφαρμογές και αξιολογήσεις των διάφορων τεχνικών που έχουν προταθεί και είδαμε και εδώ, προς αυτήν τη κατεύθυνση. Δύο τελευταίες προσεγγίσεις που λαμβάνουν υπόψη το μέτρο του πληροφοριακού περιεχομένου είναι οι (Hariri, Abolhassani, & Khodaei, 2006), (Jiang & Conrath, 1997). Οι βασικές όμως, που στάθηκαν σαν αφετηρία ήταν αυτές των Lin και Formica που είδαμε παραπάνω.

²³ ISA ιεραρχίες

Κεφάλαιο 2 – Το Πρόγραμμα DBE και το Τεχνικό Υπόβαθρο

2.1. Το ερευνητικό Πρόγραμμα DBE

Ο συνολικός στόχος του DBE ήταν, να παρέχει στην Ευρώπη ένα συγκριτικό πλεονέκτημα για την ανάπτυξη καινοτόμων εφαρμογών (λογισμικού) από την βιομηχανία των μικρομεσαίων επιχειρήσεων.

Ο J.F. Moore (Moore, 1996) περιγράφει ένα επιχειρηματικό οικοσύστημα ως: «μια οικονομική κοινότητα υποστηριζόμενη από ένα σύνολο οργανώσεων και ατόμων – τους οργανισμούς του επιχειρηματικού κόσμου. Αυτή η κοινότητα παράγει αγαθά και υπηρεσίες με αξία για τους πελάτες, οι οποίοι με τη σειρά τους είναι μέλη του οικοσυστήματος. Με την πάροδο του χρόνου εξελίσσουν παράλληλα τις δυνατότητες και τους ρόλους τους ενώ τείνουν να ευθυγραμμίζουν τους εαυτούς τους με τις μελλοντικές κατευθύνσεις...»

Πολλές βιομηχανίες σήμερα συμπεριφέρονται σαν ένα μαζικά διασυνδεδεμένο δίκτυο από οργανώσεις, τεχνολογίες, καταναλωτές και προϊόντα. Σε αντίθεση με τα «κάθετης²⁴» διάρθρωσης ολοκληρωμένα περιβάλλοντα των δεκαετιών 1960 και 1970, οι σημερινές βιομηχανίες χωρίζονται σε ένα μεγάλο αριθμό από τμήματα που παράγουν εξειδικευμένα προϊόντα, τεχνολογίες και υπηρεσίες. Ο βαθμός της αλληλεπίδρασης μεταξύ των εταιριών στη βιομηχανία είναι πραγματικά εκπληκτικός και η επιτυχία μιας επιχείρησης σήμερα δεν εξαρτάται πια από το μέγεθος των εσωτερικών και εξωτερικών της εργασιών αλλά από τις σχέσεις που καταφέρνει αυτή να αναπτύξει μέσα στο επιχειρηματικό οικοσύστημα. Το «Ψηφιακό Επιχειρηματικό Οικοσύστημα»(Digital Business Ecosystem) θέλει να αποτελέσει την ικανή τεχνολογία για το επιχειρηματικό οικοσύστημα. Ένα ψηφιακό επιχειρηματικό οικοσύστημα ορίζεται σαν ένα «εξελικτικό σύστημα με την ικανότητα αυτό-οργάνωσης στοχεύοντας στη δημιουργία ενός ψηφιακού λογισμικού περιβάλλοντος για μικρούς οργανισμούς» που θα υποστηρίζει την ανάπτυξη σε περιφερειακό και τοπικό επίπεδο. Αυτό θα επιτευχθεί δίνοντας

²⁴ Ο όρος «κάθετης» διάρθρωσης αναφέρεται στα περιβάλλοντα τα οποία ενσωματώνουν τα ίδια, τα διάφορα στάδια της παραγωγής αντί της ανάθεσής τους σε εξωτερικές, για αυτόν τον σκοπό οργανώσεις. (vertically integrated)

έμφαση σε ανοιχτές, κατανεμημένες και προσαρμόσιμες τεχνολογίες και εξελικτικά επιχειρηματικά μοντέλα για την ανάπτυξη των μικρών οργανισμών.

Το πρόγραμμα ονομάστηκε «Ψηφιακό Επιχειρηματικό Οικοσύστημα»(**D**igital **B**usiness **E**cosystem) με σκοπό να δοθεί έμφαση σε ότι θα είχε να προσφέρει σε αυτό, η παρατήρηση της λειτουργίας των βιολογικών συστημάτων. Στην πραγματικότητα το όραμα του DBE ήταν βασισμένο σε δύο βασικές ιδέες: αυτή της αυτό-οργάνωσης και αυτή της βιολογικής εξέλιξης. Η έννοια της αυτό-οργάνωσης υπονοεί την έξυπνη συμπεριφορά του συστήματος και την ικανότητά του να μαθαίνει σε σύντομο χρονικό διάστημα. Ενώ η έννοια της εξέλιξης υπονοεί την ικανότητά του να βελτιώνεται μέσω της μακροπρόθεσμης διαφοροποίησης και επιλογής των χαρακτηριστικών του.

Μετά τα παραπάνω ο όρος DBE μπορεί να γίνει αντιληπτός και να βρει εφαρμογή στα εξής επίπεδα:

1. τα ανοιχτού κώδικα κατανεμημένα περιβάλλοντα (υποδομή και στοιχεία ενδιάμεσου λογισμικού²⁵) μέσα στα οποία τμήματα και υπηρεσίες αλληλεπιδρούν και εξελίσσονται
2. το περιβάλλον, μαζί με τα τμήματά του και τις υπηρεσίες που επιτρέπουν την ανακάλυψη²⁶ και την εξέλιξη, δηλαδή το ζωτικό επίπεδο του συστήματος
3. τις εφαρμογές τις οποίες τελικά χρησιμοποιούν οι επιχειρήσεις
4. και τελικά, σε ποιο υψηλό επίπεδο, τα επιχειρηματικά μοντέλα εξέλιξης όπως αυτά επηρεάζονται από το αναπτυσσόμενο λογισμικό και την αλληλεπίδραση μεταξύ των ίδιων των επιχειρήσεων.

Στην πραγματικότητα κανείς, μπορεί να αντιληφθεί το DBE σαν μια ανάμειξη των παραπάνω επιπέδων σε ένα εξελικτικό σύστημα με την ικανότητα αυτό-οργάνωσης και στόχο τη δημιουργία ενός λογισμικού περιβάλλοντος για μικρομεσαίες επιχειρήσεις.

Ο στόχος ήταν τα αποτελέσματα του DBE να αποτελέσουν μια στροφή στην πρόσβαση, την διανομή και την σύνθεση λογισμικού ή οποία θα αλλάξει τον τρόπο

²⁵ Ενδιάμεσο λογισμικό εννοώντας το λογισμικό που μεσολαβεί μεταξύ δύο διαφορετικών εφαρμογών οι οποίων οι στοιχείων λογισμικού, συνήθως για την επικοινωνία τους (middleware).

²⁶ Η ανακάλυψη των υπολοίπων μελών και παρεχόμενων υπηρεσιών του οικοσυστήματος είναι ζωτικής σημασίας για την ανάπτυξη και εξέλιξη του συστήματος. Χρειάζονται υπηρεσίες που να υποστηρίζουν αυτή την αναζήτηση και ανακάλυψη και φυσικά αυτή να είναι εφικτή από την υποδομή του συστήματος.

με τον οποίο οι μικρομεσαίες επιχειρήσεις και οι διανομείς λογισμικού θα χρησιμοποιούν και θα διανέμουν τα προϊόντα και τις υπηρεσίες τους.

Οι τέσσερις βασικές απαιτήσεις του DBE ήταν οι ακόλουθες:

- Η ικανότητα να παράγει δομές λογισμικού που θα ανταποκρίνονται στις απαιτήσεις των μικρομεσαίων επιχειρήσεων
- Η ικανότητα να εξελίσσει και να προσαρμόζει το λογισμικό καθώς αλλάζουν οι απαιτήσεις με το τη πάροδο του χρόνου
- Η ικανότητα να παράγει βελτιούμενες δομές λογισμικού πέρα από τις ελάχιστες απαιτήσεις των μικρομεσαίων επιχειρήσεων, επηρεάζοντας με αυτόν τον τρόπο θετικά τα επιχειρηματικά τους μοντέλα
- Η ικανότητα να προσαρμόζει την δομή του στα τοπικά οικοσυστήματα.

Όλα τα παραπάνω μπορεί να δίνουν μια αρκετά γενική περιγραφή δεν είναι όμως μακριά από την πραγματική φύση του προγράμματος. Αναφερόμαστε στην διεπιστημονικότητα που αυτό απαιτούσε προκειμένου να επιτευχθούν οι παραπάνω στόχοι και τελικά να υποστηριχθούν και σε τεχνικό επίπεδο. Ακριβώς μέσα από τους επιστημονικούς στόχους του προγράμματος θα προσπαθήσουμε να γίνουμε πιο συγκεκριμένοι στην επόμενη παράγραφο πριν περάσουμε στις πιο τεχνικές λεπτομέρειες που αφορούν και την παρούσα εργασία.

2.1.1. Επιστημονικοί Στόχοι

Η ερευνητική πρόκληση του DBE απλώνεται σε τρεις πρωταρχικές περιοχές: την επιχειρηματική, την επιστημονική και της επιστήμης υπολογιστών. Σε μια πιο ολιστική προσέγγιση το DBE απαιτεί βασική έρευνα στην αυτό-οργάνωση σύνθετων συστημάτων, στις κοινωνικές επιστήμες και στις αρχιτεκτονικές δικτύων. Σε αυτό το σημείο γίνεται ορατή η διεπιστημονική προσέγγιση που απαιτείται για την σχεδίαση, δημιουργία και υλοποίηση του DBE.

Η κύρια επιστημονική εργασία αφορά την μετάφραση της έννοιας που αναπτύχθηκε για να περικλείσει τις κοινωνικές οργανώσεις και τους ζωντανούς οργανισμούς, σε ένα σύνολο από κατάλληλες έννοιες και λειτουργικά μοντέλα για την ανάπτυξη του ψηφιακού λογισμικού οικοσυστήματος – απευθυνόμενο σε μικρές επιχειρήσεις. Η θεωρητική προσέγγιση ήταν να δημιουργηθεί ένα

εννοιολογικό πλαίσιο για την ανάπτυξη ενός πραγματικού τεχνολογικού δικτύου που θα περιλαμβάνει θεωρία σύνθετων συστημάτων, στατιστικούς μηχανισμούς μαθηματικής μοντελοποίησης, και προχωρημένη έρευνα σε γενετικούς και εξελικτικούς αλγορίθμους. Τα παραπάνω θα πρέπει να «παντρευτούν» με τη μηχανική και τις επιστήμες των υπολογιστών κατά την συναρμολόγηση του οικοσυστήματος.

Στο υπόλοιπο του κεφαλαίου θα ασχοληθούμε με τους στόχους που αφορούσαν τους δικούς μας τομείς, της επιστήμης υπολογιστών, και ακόμα πιο συγκεκριμένα την ανάπτυξη του τεχνολογικού δικτύου, όπως αναφέρθηκε στην προηγούμενη παράγραφο.

2.1.2. Η συνεισφορά στο DBE

Η συνεισφορά μας στο DBE περιλάμβανε κυρίως την υλοποίηση της αρχιτεκτονικής του συστήματος, την προσφορά εργαλείων «φιλικών» στον χρήστη που θα επέτρεπαν τον ορισμό υπηρεσιών και πολιτικών χρησιμοποιώντας τις γλώσσες του DBE και τέλος την παραγωγή, εγκατάσταση, χρήση και πλήρωση του DBE με μικρομεσαίες επιχειρήσεις, υπηρεσίες και σχέσεις.

Από τεχνικής άποψης η αρχιτεκτονική του DBE παρέχει μια δυναμική, αυτό-διαχειριζόμενη, ανοικτού κώδικα, κατανεμημένη πλατφόρμα η οποία υποστηρίζει την εγκατάσταση, ανάκτηση και σύνθεση υπηρεσιών λογισμικού. Η πλατφόρμα κτίστηκε πάνω σε ένα ομότιμο²⁷ δίκτυο κατανεμημένης αρχιτεκτονικής παρακάμπτοντας με αυτόν τον τρόπο την ανάγκη για κεντρική αποθήκευση, διαχείριση και έλεγχο των υπηρεσιών και των δεδομένων.

Το DBE βασίστηκε πάνω σε ένα υπερκείμενο ομότιμο δίκτυο με επιπλέον υποστήριξη για πολλαπλά αλληλεπιδραστικά μοντέλα μεταξύ των υπηρεσιών όπως:

ενδιάμεσο λογισμικό μηνυμάτων (message-oriented middleware - MOM)

απομακρυσμένη κλήση διαδικασιών (remote procedure call - RPC)

ενδιάμεσο λογισμικό συνδιαλλαγών (transaction-oriented middleware)

²⁷ Ομότιμο δίκτυο, δηλαδή δίκτυο ομότιμων κόμβων(peer-to-peer)

Το ενδιαμέσο λογισμικό ήταν μια επέκταση της έκδοσης 2 του FADA²⁸ υλοποιημένο από την Sun Microsystems, συνεργάτη του προγράμματος DBE. Για το FADA θα μιλήσουμε στην [§2.3.2] παράγραφο αναφερόμενοι στα τεχνικά χαρακτηριστικά της υλοποίησης.

Το ενδιαμέσο λογισμικό ομότιμων κόμβων υπόσχεται, ως γνωστόν, βελτιωμένες δυνατότητες ως προς το μέγεθος και την απόδοση του συστήματος συγκριτικά με τα παραδοσιακά συστήματα πελάτη-εξυπηρετητή, καθώς γίνεται καλύτερη εκμετάλλευση του δικτύου και των υπολογιστικών πόρων. Τεχνικές όπως οι κατανεμημένοι πίνακες γρήγορης δεικτοδότησης (DHTs)²⁹, τα μοντέλα περιορισμένης εκπομπής (broadcast models), η πιθανοτική δρομολόγηση (probabilistic routing), τα πλέγματα υπολογιστών και τα μοντέλα πλεονασμών (redundancy models) καθιστούν τις συγκεντρωτικές ενδιαμέσες υπηρεσίες περιττές. Με λίγα λόγια σήμερα υπάρχουν οι τεχνικές και η υποδομή λογισμικού για την αποφυγή συστημάτων με συγκεντρωτικό έλεγχο και διαχείριση. Από την άλλη πλευρά, η κατανομή των κοινών μετα-δεδομένων του DBE στους διάφορους κόμβους επιτρέπει μια πιο αποδοτική χρήση των πόρων και την αποφυγή ακριβώς αυτής της συγκεντρωτικής διαχείρισης του συστήματος. Παραδείγματα κοινών μετα-δεδομένων περιλαμβάνουν δεδομένα εκκίνησης, δεδομένα ελέγχου, στατιστικά δεδομένα, προφίλ υπηρεσιών, διαθεσιμότητας κοινών πόρων κ.ά.

Τα παραπάνω αποτέλεσαν από την αρχή τον οδηγό για τις απαιτήσεις κατά το σχεδιασμό του συστήματος και είχαν σαν αποτέλεσμα την υιοθέτηση ενός συγκεκριμένου τεχνικού πλαισίου που θα υποστήριζε την επιθυμητή συμπεριφορά και χαρακτηριστικά του συστήματος. Το τεχνικό αυτό πλαίσιο θα περιγράψουμε στο υπόλοιπο του κεφαλαίου δίνοντας έμφαση στα κομμάτια που αποτέλεσαν αντικείμενο της παρούσας εργασίας.

2.2. Τεχνικό Πλαίσιο και Τεχνολογίες

2.2.1. Το δίκτυο

²⁸ Πλατφόρμα υλοποίησης για την υποστήριξη ομότιμων δικτύων σε φυσικό επίπεδο (Sourceforge.net)

²⁹ Γρήγορη δεικτοδότηση με χρήση hashing. Αναφερόμαστε σε distributed hash tables σαν μια από τις κλασικές τεχνικές αντιμετώπισης κατανεμημένων δικτύων.

Όπως έχει ήδη παρουσιαστεί από τις προηγούμενες παραγράφους η αρχιτεκτονική του DBE βασίστηκε σε ένα ομότιμο δίκτυο. Τα πλεονεκτήματα των ομότιμων δικτύων είναι γνωστά και τα ίδια αποτελούν μια πολύ διαδεδομένη αλλά και δοκιμασμένη πρακτική σε πολλά σημερινά συστήματα. Οι λόγοι που υιοθετήθηκε ένα ομότιμο δίκτυο για τις ανάγκες του DBE ήταν οι εξής:

Η υλοποίηση του δικτύου σε φυσικό επίπεδο δεν ήταν στους στόχους του προγράμματος δεδομένου μάλιστα ότι σήμερα υπάρχουν μερικές καλές πλατφόρμες υλοποίησης ομότιμων δικτύων όπως το JXTA κ.α. Στη περίπτωση του DBE αποφασίστηκε να χρησιμοποιηθεί η πλατφόρμα FADA. Έχοντας έτοιμη τη πλατφόρμα FADA έπρεπε να γίνει προσαρμογή της στις ανάγκες του προγράμματος και εγκατάστασή της σε πραγματικό περιβάλλον. Πιο συγκεκριμένα χρειάστηκε να γίνουν επεμβάσεις στον πυρήνα του FADA ώστε να ενσωματωθούν όλες οι υπηρεσίες που θα ήταν μελλοντικά διαθέσιμες στους χρήστες του δικτύου. Με αυτόν τον τρόπο κάθε κόμβος του δικτύου μαζί με την εγκατάσταση του λογισμικού που τον έκανε μέλος του δικτύου του DBE αποκτούσε και μια σειρά από υπηρεσίες διαθέσιμες τόσο στον ίδιο όσο και στους πελάτες του. Καθώς το δίκτυο ήταν ομότιμο το ίδιο ίσχυε και για όλα τα υπόλοιπα μέλη του δικτύου με αποτέλεσμα να έχουν όλοι ταυτόχρονα το ρόλο του πελάτη αλλά και του εξυπηρετητή γεγονός που αποτυπώθηκε εξ' αρχής στο λογισμικό των κόμβων αποδίδοντάς του το όνομα Υπηρέτης-Πελάτης (ServEnt³⁰). Στα πλαίσια και της παρούσας εργασίας μια σειρά από υπηρεσίες υλοποιήθηκαν με σκοπό τη δημοσίευση συγκεκριμένης λειτουργικότητας.

2.2.2. Η Βάση Γνώσης

Όσα περιγράφονται στην προηγούμενη παράγραφο αποτέλεσαν το φυσικό υπόβαθρο για τη δημιουργία του δικτύου ενώ οι υπηρεσίες που δημιουργήθηκαν υποστήριζαν την επικοινωνία μεταξύ των κόμβων για τις ανάγκες και με τη λογική του DBE. Το επόμενο και απαραίτητο βήμα ήταν η εύρεση ενός κοινού τρόπου περιγραφής, αποθήκευσης και διαχείρισης των δεδομένων που χρειάζεται να

³⁰ Προέκυψε από σύντμηση των λέξεων Server και Client.

ανταλλάσσονται και να διακινούνται. Αυτό όπως αναφέρθηκε και στο 1 Κεφάλαιο ήταν το πλαίσιο αναπαράστασης της γνώσης(Knowledge Representation Framework) του DBE. Η όλη προσέγγιση για την μοντελοποίηση και αναπαράσταση της γνώσης έπρεπε να καλύπτει τις λειτουργικές και τεχνικές απαιτήσεις του DBE. Για το σκοπό αυτό όλη η απαιτούμενη γνώση βρίσκεται στη Βάση Γνώσης του DBE. Για να υπάρχει σημασιολογική διαλειτουργικότητα κρίσιμος είναι ο ρόλος των οντολογιών, ένα εργαλείο που εξ' ορισμού σήμερα αναλαμβάνει αυτόν το ρόλο. Ένα μοντέλο ορισμού οντολογιών(Ontology Definition Metamodel) αναπτύχθηκε από τη βάση του, στο DBE. Παράλληλα αναπτύχθηκαν μηχανισμοί για ορισμό τόσο επιχειρηματικών μοντέλων όσο και οντολογιών για τις διάφορες υπηρεσίες.

Το πλαίσιο αναπαράστασης της γνώσης ακολουθεί την αρχιτεκτονική μεταδεδομένων MOF του OMG³¹ όπου ορίζονται τέσσερα επίπεδα αφαίρεσης για την μοντελοποίηση, αναπαράσταση και οργάνωση των δεδομένων και των μεταδεδομένων. Με μια φράση, σύμφωνα με τη MOF αρχιτεκτονική κάθε κομμάτι πληροφορίας που βρίσκεται αποθηκευμένο στη βάση γνώσης, έχει παραχθεί και τοποθετηθεί σαν ένα στιγμιότυπο κάποιου μοντέλου υψηλότερου επιπέδου. Ο πυρήνας του πλαισίου αναπαράστασης γνώσης στο DBE αποτελείται από οντολογίες διαφορετικών πεδίων. Αυτές οι οντολογίες αποτυπώνουν όλη την κοινώς αποδεκτή σημασιολογία (έννοιες και σχέσεις μεταξύ τους) ενός συγκεκριμένου επιχειρηματικού πεδίου και αποτελούν το σημείο αναφοράς για κάθε συγκεκριμένο μοντέλο ή περιγραφή επιτρέποντας από τη μια πλευρά το διαμοιρασμό της πραγματικής γνώσης και από την άλλη πλευρά την σημασιολογική διαλειτουργικότητα μεταξύ των μικρομεσαίων επιχειρήσεων. Με λίγα λόγια με αυτόν τον τρόπο, οι μικρομεσαίες επιχειρήσεις μοντελοποιούν και περιγράφουν την επιχείρησή τους αλλά και της προσφερόμενες υπηρεσίες τους με κοινά αποδεκτές και κατανοητές έννοιες από όλη την κοινότητα. Η αναπαράσταση οντολογιών συγκεκριμένων πεδίων βασισμένη στη MOF αρχιτεκτονική οδήγησε στην ανάγκη για ένα νέο μεταμοντέλο. Ένα μεταμοντέλο που να μπορεί από τη μια πλευρά να περιγράφει καινούριες οντολογίες και από την άλλη πλευρά να αναπαριστά υπάρχουσες οντολογίες μετατρέποντας το DBE σε ένα ανοιχτό περιβάλλον. Την

³¹ Object Management Group, το επίσημο γκρουπ για την τυποποίηση, οργάνωση και διαχείριση μέσω προτύπων στο χώρο διαχείρισης αντικειμένων γενικά www.omg.org

ανάγκη αυτή αναγνώρισε και το ίδιο το OMG δημοσιεύοντας σχετικό κάλεσμα για προτάσεις (OMG, The Object Management Group, ODM RFP). Το DBE ανταποκρίθηκε στο παραπάνω κάλεσμα δημιουργώντας και προτείνοντας το μεταμοντέλο ορισμού οντολογιών ODM (Ontology Definition Metamodel). Πέρα από αυτό υπήρχε και η ανάγκη περιγραφής συγκεκριμένων επιχειρηματικών υπηρεσιών και σχετικών οντολογιών καθώς και μηχανισμών που θα επέβαλλαν την χρήση συγκεκριμένων εννοιών και σημασιολογίας κατά τη δημιουργία των μοντέλων. Για τον λόγο αυτό ακολούθησε ο ορισμός εξειδικευμένων γλωσσών μοντελοποίησης όπως ήταν η Γλώσσα Σημασιολογικών Υπηρεσιών (Semantic Service Language) και η Γλώσσα Μοντελοποίησης Επιχείρησης (Business Modeling Language). Δεν θα χρειαστεί να αναφερθούμε σε περισσότερες λεπτομέρειες για αυτές τις γλώσσες. Το σημαντικό στοιχείο είναι ότι υπάρχει ένα ολοκληρωμένο πλαίσιο για την περιγραφή και αποθήκευση της γνώσης βασισμένο στο MOF και την χρήση οντολογιών. Για το MOF θα αναφερθούμε λίγο περισσότερο στο τέλος του κεφαλαίου ενώ για τις οντολογίες θα αναφερθούμε εκτενέστερα επίσης στο τέλος του κεφαλαίου καθώς αποτελούν το κεντρικό άξονα της παρούσας εργασίας. Αυτό που είναι σημαντικό να σημειώσουμε σε αυτό το σημείο είναι ότι όλη η υποστήριξη από πλευράς οντολογιών είναι συνεπής με την γλώσσα περιγραφής οντολογιών OWL (Bechhofer, και συν., 2004).

Εκτός από τις περιγραφές και την αποθήκευση της γνώσης χρειαζόταν φυσικά και ένα μηχανισμό ανάκτησής της. Ο μηχανισμός αυτός όπως αναφέρθηκε και στο 1 κεφάλαιο στηρίχθηκε πάνω σε μια καινούρια γλώσσα πρόσβασης στη γνώση του DBE, βασισμένη στο MOF, την Γλώσσα Ερωτήσεων Μεταμοντέλων – QML (Query Metamodeling Language (Kotopoulos, Kazasis, & Christodoulakis, 2007)). Στην QML θα αναφερθούμε αναλυτικότερα στο επόμενο κεφάλαιο [§2.5].

2.3. Περίγραμμα Εργασίας

Μέχρι στιγμής περιγράψαμε τους στόχους του ερευνητικού προγράμματος DBE. Επίσης αναφερθήκαμε στο τεχνικό και στο θεωρητικό πλαίσιο του προγράμματος και αυτό ήταν και το γενικότερο πλαίσιο της παρούσας εργασίας. Στο υπόλοιπο του κεφαλαίου θα επικεντρωθούμε και θα αναλύσουμε περισσότερο τα κομμάτια εκείνα που είχαν άμεση σχέση με ότι υλοποιήθηκε και προτάθηκε κατά τη διάρκεια της εργασίας. Προκειμένου να υπάρχει μια καλύτερη εικόνα για τις κατευθύνσεις

στις οποίες κινηθήκαμε θα θέλαμε να θυμίσουμε το περίγραμμα της εργασίας όπως αυτό δόθηκε στην παράγραφο [§1.2]. Πιο συγκεκριμένα σχετικά με το ίδιο το DBE και το τεχνικό κομμάτι χρειάστηκε:

- Εξοικείωση, τροποποίηση, προσαρμογή της έτοιμης πλατφόρμας ανοιχτού κώδικα FADA
- Ανάπτυξη ενός πλαισίου αποθήκευσης, διαχείρισης και ανάκτησης των δεδομένων και των μεταδεδομένων του συστήματος
- Δημιουργία μιας μεθοδολογίας ερωτήσεων και ανάκτησης δεδομένων τόσο σε τοπικό επίπεδο όσο και σε επίπεδο δικτύου

Στο τρέχον κεφάλαιο θα ασχοληθούμε με τα δύο πρώτα και στο τέλος θα δούμε τη γλώσσα ερωτήσεων που χρησιμοποιήθηκε ενώ το υπόλοιπο κομμάτι των ερωτήσεων θα αποτελέσει αντικείμενο του επόμενου κεφαλαίου.

2.3.1. Φυσικό Δίκτυο

Στη παράγραφο αυτή θα παρουσιάσουμε τα στοιχεία που σχετίζονται με την ανάπτυξη του φυσικού δικτύου του συστήματος. Επικεντρωνόμαστε σε όσα από αυτά αποτέλεσαν άμεσα τμήμα της παρούσας εργασίας. Πιο συγκεκριμένα, περιληπτικά, για το φυσικό δίκτυο απαιτήθηκε εγκατάσταση της πλατφόρμας FADA για την οποία θα αναφερθούμε στα βασικά της χαρακτηριστικά. Στη συνέχεια χρειάστηκε εξοικείωση με την ίδια την υλοποίηση της πλατφόρμας προκειμένου να αναπτυχθούν στον πυρήνα της νέα κομμάτια λειτουργικότητας που θα εξυπηρετούσαν τις ανάγκες της εργασίας. Στα κομμάτια αυτά επίσης θα αναφερθούμε στις επόμενες παραγράφους. Τέλος έγινε εγκατάσταση του δικτύου σε πραγματικό περιβάλλον τόσο στο χώρο του εργαστηρίου Διανεμημένων Πληροφοριακών Συστημάτων και Εφαρμογών όσο και σε απομακρυσμένο επίπεδο, στις περιοχές των συνεργατών του προγράμματος DBE. Στην αμέσως επόμενη παράγραφο ξεκινάμε την περιγραφή μας με τα χαρακτηριστικά του δικτύου.

Στην αρχή του προγράμματος υπήρχαν συγκεκριμένες προτάσεις για την ανάπτυξη ενός μερικά διασυνδεδεμένου δικτύου. Παρά τα κάποια πλεονεκτήματα ενός τέτοιου δικτύου, σύντομα το γεγονός ότι οι μικρομεσαίες επιχειρήσεις βρίσκονται σε γεωγραφικά απομακρυσμένες περιοχές και δεν θα μπορούσαν να συνδέονται αυτόματα στο δίκτυο είτε θα έπρεπε να έχουν πρόσβαση μόνο μέσω άλλων

«ισχυρά» συνδεδεμένων επιχειρήσεων οδήγησε την υιοθέτηση ενός πλήρως συνδεδεμένου δικτύου. Σε αυτές τις περιπτώσεις ο νόμος του Metcalfe, ότι η χρησιμότητα ενός δικτύου είναι ανάλογη του τετραγώνου του μεγέθους του δικτύου, παίζει αποφασιστικό ρόλο. Το επόμενο χαρακτηριστικό του δικτύου ήταν ότι θα έπρεπε να ενσωματώνει τη δυνατότητα για κάποιες υπηρεσίες με σημαντικότερη αυτή της αναζήτησης και μάλιστα της σημασιολογικής αναζήτησης, απαίτηση που καλύφθηκε με την δημιουργία ενός σημασιολογικού μητρώου(Semantic Registry). Παράλληλα υλοποιήθηκαν και ενσωματώθηκαν μια σειρά από άλλες υπηρεσίες και στα πλαίσια της παρούσας εργασίας, τις οποίες θα δούμε σε επόμενη παράγραφο. Στα χαρακτηριστικά του δικτύου έπρεπε να ληφθούν υπόψη και στοιχεία όπως οι «τοίχοι» ασφαλείας ή η φυσική ετερογένεια των κόμβων του δικτύου, κάτι όμως που ήταν έξω από τους σκοπούς μας.

Σε αυτό το σημείο και μιλώντας για το φυσικό δίκτυο, είναι σημαντικό να αναφέρουμε ότι το σύστημα στηρίχθηκε σε μια αρχιτεκτονική η οποία περιελάμβανε μεν ένα ομότιμο δίκτυο παράλληλα όμως ήταν δομημένη σε διαφορετικά επίπεδα δικτύου. Πιο συγκεκριμένα όσα περιγράφονται στην παρούσα παράγραφο αφορούν το ομότιμο δίκτυο σε φυσικό επίπεδο επιτρέποντας την φυσική σύνδεση απομακρυσμένων κόμβων μεταξύ τους. Θα έπρεπε δηλαδή να μπορεί οποιοσδήποτε κόμβος και οποιαδήποτε στιγμή να γίνει μέλος του δικτύου και αυτομάτως να γίνεται γνωστή η φυσική του παρουσία. Αυτός ήταν ακριβώς και ο ρόλος που ανέλαβε το FADA επιτρέποντας κατά τον ίδιο τρόπο και την ανακάλυψη από οποιονδήποτε κόμβο των φυσικών του γειτόνων. Επιμένουμε να αναφερόμαστε στο δίκτυο σε αυτό το σημείο ως «φυσικό» θέλοντας να το διαχωρίσουμε από το άλλο επίπεδο δικτύου, το υπερκείμενο δίκτυο(Overlay Network). Πρόκειται για το δίκτυο σε πιο υψηλό επίπεδο όπου οι συνδέσεις μεταξύ των μελών του πραγματοποιούνται όχι με βάση τη φυσική δυνατότητα σύνδεσης μεταξύ τους αλλά αξιοποιώντας κάποιο πρωτόκολλο σύνδεσης με βάση για παράδειγμα σημασιολογικά κριτήρια. Για τον ίδιο λόγο ανάλογα με τα κριτήρια σύνδεσης των κόμβων μεταξύ τους είναι δυνατόν να συνυπάρχουν περισσότερα του ενός υπερκείμενων δικτύων. Όλες οι προτάσεις και η συνεισφορά της παρούσας εργασίας αναφέρονται ακριβώς σε αυτό το υπερκείμενο δίκτυο έχοντας ουσιαστικά την σημασιολογική οπτική του συστήματος. Πριν περάσουμε όμως σε αυτά θα κλείσουμε την τρέχουσα παράγραφο με λίγα λόγια για το FADA.

2.3.2. Το FADA (Federated Autonomous Directory Architecture)

Το FADA είναι πρακτικά ένας διανεμημένος κατάλογος από υπηρεσίες. Πρόκειται για έναν εικονικό εξυπηρετητή αναζήτησης αυτών των υπηρεσιών με την έννοια ότι διαφορετικοί εξυπηρετητές αναζήτησης συνεργάζονται για να παρέχουν την υπηρεσία αναζήτησης από οποιοδήποτε σημείο του δικτύου. Το FADA είναι ένα πραγματικό διανεμημένο σύστημα καθώς δεν υπάρχει καμία κεντρική αρχή ή κοινό κανάλι επικοινωνίας και όλοι οι εξυπηρετητές αναζήτησης έχουν τον χαρακτήρα ομότιμων κόμβων. Το σύστημα διαθέτει εξουσιοδότες(proxyes) για τις διάφορες υπηρεσίες. Κάθε εξουσιοδότης (proxy) είναι μια κλάση σε Java, η οποία πραγματοποιεί την επικοινωνία με μια προσφερόμενη υπηρεσία και η οποία μπορεί να αποκτηθεί κατά την εκτέλεση από οποιονδήποτε πελάτη. Οι πελάτες χρησιμοποιούν τις δημόσιες μεθόδους των εξουσιοδοτών για να έχουν πρόσβαση τις υπηρεσίες. Αυτές οι δημόσιες μέθοδοι είναι ορισμένες σαν Java διεπαφές και υλοποιούνται από τους εξουσιοδότες. Κάθε κόμβος του FADA είναι μια υπηρεσία που λειτουργεί σαν ένα σημείο εισόδου στη Βάση Γνώσης. Οποιοσδήποτε θέλει να γίνει παροχέας μιας υπηρεσίας θα πρέπει να υλοποιεί την αντίστοιχη κλάση που θα είναι η πύλη για την υπηρεσία του.

Το FADA χρησιμοποιεί την υποδομή του DNS, ένα γνωστό όνομα περιοχής, και την ικανότητα των κόμβων του FADA να γνωρίζουν την γειτονιά τους, για να παρέχει την δυνατότητα ανακάλυψης νέων κόμβων καθώς και της τοποθεσίας τους. Σαν τοποθεσία αναφερόμαστε σε καθαρά φυσικό επίπεδο στην IP διεύθυνση στην οποία ο κόμβος αναμένει για αιτήσεις. Με αυτόν τον τρόπο χτίζεται ένα δίκτυο από συνεργαζόμενους FADA κόμβους οι οποίοι χρησιμοποιούν απευθείας http συνδέσεις μεταξύ τους.

Σαν τοπολογία δικτύου επιλέχτηκε η τυχαία τοπολογία³². Δεν γίνεται δηλαδή η επιβολή κάποιας συγκεκριμένης δομής στο δίκτυο. Κάθε κόμβος μπορεί να συνδεθεί με οποιονδήποτε άλλο κόμβο και η τοπολογία είναι δυναμική και όχι γνωστή εκ των προτέρων. Αυτό δίνει μεγάλη ευελιξία στο δίκτυο: κάθε κόμβος μπορεί οποιαδήποτε στιγμή να γίνει μέλος και σε οποιοδήποτε σημείο και αυτομάτως να προσπελαστεί από οποιονδήποτε πελάτη το επιθυμήσει. Επιπρόσθετα ένας κόμβος

³² Random network topology

μπορεί οποιαδήποτε στιγμή να φύγει και οι υπόλοιποι να συνεχίσουν τη λειτουργία χωρίς κανένα πρόβλημα ή αλλαγή.

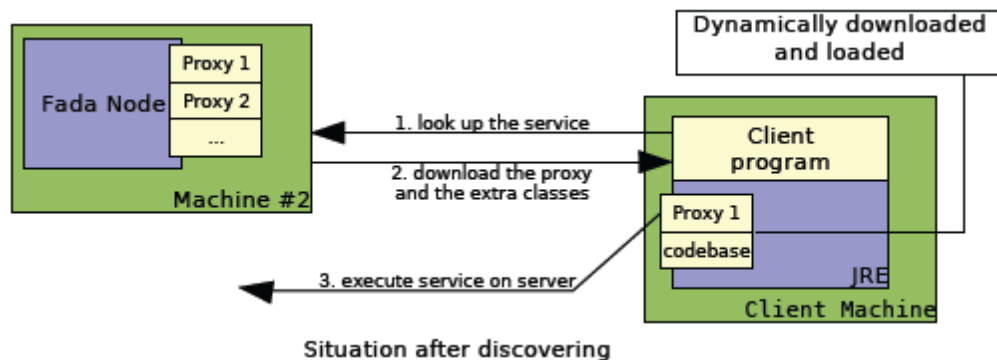
Συνοψίζοντας τα παραπάνω οι μηχανισμοί που χρησιμοποιούνται από το FADA εξυπηρετούν δύο στόχους: την δυνατότητα αποθήκευσης και ανάκτησης εξουσιοδοτών υπηρεσιών και τη δυνατότητα διαχείρισης της δομής ενός γράφου από FADA κόμβους. Από την στιγμή που ένας κόμβος εγκαθίσταται μπορεί να χρησιμοποιηθεί για την αποθήκευση εξουσιοδοτών υπηρεσιών. Οι παροχείς υπηρεσιών πρέπει να προσφέρουν έναν εξουσιοδότη για κάθε μια από τις υπηρεσίες τους. Αυτός ο εξουσιοδότης πρέπει να είναι γραμμένος σε Java αλλά ο παροχέας μπορεί να χρησιμοποιήσει οποιαδήποτε υλοποίηση για την ίδια την υπηρεσία (Java RMI, HTTP, JNI, JXTA κ.ο.κ.), έχοντας πάντα κατά νου ότι οι εξουσιοδότες θα αποκτηθούν από τους πελάτες και θα εκτελεστούν τοπικά στην εικονική μηχανή εκτέλεσης της Java του πελάτη. Τελικά, τα βήματα για την καταχώρηση μιας υπηρεσίας σε ένα κόμβο περιλαμβάνουν:

1. Να υπάρχει μια ενεργοποιημένη υπηρεσία. Είναι αναγκαία η διαφοροποίηση μεταξύ υπηρεσίας και εξουσιοδότη. Η υπηρεσία μπορεί να βρίσκεται σε οποιοδήποτε μηχανήμα και να είναι υλοποιημένη σε οποιαδήποτε γλώσσα ή τεχνολογία. Ο εξουσιοδότης είναι ουσιαστικά ένα μικρό κομμάτι κώδικα το οποίο γνωρίζει τον τρόπο επικοινωνίας με την υπηρεσία.
2. Να οριστεί η διεπαφή μιας υπηρεσίας. Η διεπαφή θα ορίσει της μεθόδους τις οποίες ο πελάτης μπορεί να εκτελέσει για την επικοινωνία του με την υπηρεσία. Αυτή η διεπαφή είναι προσβάσιμη από όλους του πελάτες και είναι γραμμένη σε Java.
3. Να υπάρχει ένα εξουσιοδότης υπηρεσίας. Ο εξουσιοδότης είναι πρακτικά η υλοποίηση της διεπαφής μιας υπηρεσίας. Η κλάση του εξουσιοδότη εκτελείται στην πλευρά του πελάτη και είναι επίσης γραμμένη σε Java.
4. Να καταχωρηθεί ο εξουσιοδότης σε έναν FADA κόμβο. Μόνο τότε γίνεται ο εξουσιοδότης διαθέσιμος και μπορεί να ανακαλυφθεί από τους πελάτες του δικτύου.
5. Να είναι διαθέσιμο ένα jar αρχείο με βιβλιοθήκες. Η υλοποίηση του εξουσιοδότη δεν είναι γνωστή στους πελάτες και για αυτό θα πρέπει να αποκτηθεί από αυτούς. Ο εξυπηρετητής θα πρέπει να έχει προσβάσιμες όλες τις επιπλέον κλάσεις σε ένα jar αρχείο. Η τοποθεσία αυτού του αρχείου ονομάζεται codebase.

Η διαδικασία από την πλευρά του πελάτη ανταποκρίνεται στις παραπάνω απαιτήσεις και περιλαμβάνει τα εξής τρία βήματα:

1. Να έχει πρόσβαση στην διεπαφή της υπηρεσίας. Ο πελάτης δεν χρειάζεται να γνωρίζει κάτι για την υλοποίηση της υπηρεσίας ή του εξουσιοδότη αλλά η διεπαφή θα πρέπει να είναι προσβάσιμη.
2. Να ανακαλύψει την υπηρεσία. Για την ανακάλυψη της υπηρεσίας ο πελάτης θα πρέπει να γνωρίζει το όνομα της διεπαφής καθώς και την διεύθυνση του αντίστοιχου κόμβου του δικτύου.
3. Να αποκτήσει και να εκτελέσει τον εξουσιοδότη. Μετά την ανακάλυψη της υπηρεσίας ο εξουσιοδότης εγκαθίσταται αυτόματα και είναι έτοιμος για χρήση επιτρέποντας την απευθείας επικοινωνία με την υπηρεσία του εξυπηρετητή.

Τα παραπάνω φαίνονται και στο παρακάτω σχήμα:



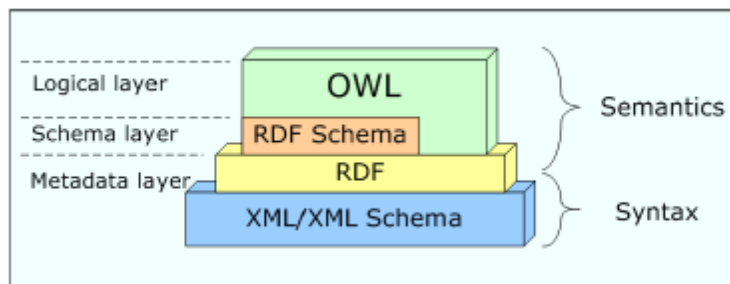
Σχήμα 2: Διαδικασία ανακάλυψης και χρήσης μιας υπηρεσίας κάποιου κόμβου του δικτύου από έναν πελάτη

2.4. Τεχνικό Πλαίσιο Εργασίας

Αφού είδαμε όλο το τεχνικό πλαίσιο για την κατασκευή ενός δικτύου με τα παραπάνω χαρακτηριστικά είναι η κατάλληλη στιγμή για να περιγράψουμε το πλαίσιο για την αποθήκευση όλων των δεδομένων και μεταδεδομένων της εφαρμογής. Τα δεδομένα αυτά όπως είχαμε δει στην παράγραφο [§1.2] αποτελούν την Βάση Γνώσης της εφαρμογής και όλο το πλαίσιο είναι το Πλαίσιο Αναπαράστασης της Γνώσης (Knowledge Representation Framework) του DBE. Στις επόμενες δύο παραγράφους θα δούμε την αρχιτεκτονική που ακολουθήθηκε για την αποθήκευση των δεδομένων της εφαρμογής μαζί με τη σημασιολογία τους

καθώς και την βασική γλώσσα της περιγραφής του που ήταν η ODM. Στην τελευταία παράγραφο του κεφαλαίου θα ασχοληθούμε με τη γλώσσα ερωτήσεων QML.

Όπως είδαμε και στο 1^ο κεφάλαιο εισαγωγικά, ο σημασιολογικός ιστός δημιουργήθηκε για να επιτρέψει τον διαμοιρασμό δεδομένων που θα είναι κατανοητά από τις μηχανές. Ο σημασιολογικός ιστός στηρίζεται στα μεταδεδομένα που περιγράφονται με τη βοήθεια οντολογιών δίνοντας έννοιες κατανοητές από τους υπολογιστές, στα δεδομένα. Αυτή η χρήση των οντολογιών γίνεται με τη βοήθεια προτύπων όπως η OWL (Bechhofer, και συν., 2004). Η OWL σήμερα αποτελεί ένα πολύ διαδεδομένου εργαλείο για την περιγραφή δεδομένων και αποτελεί και τη βάση του σημασιολογικού ιστού. Η αρχιτεκτονική του σημασιολογικού ιστού είναι μια λειτουργική, όχι σταθερή αρχιτεκτονική (Brickley & Guha, 2000). Ο Berners-Lee (Berners-Lee, 1998) όρισε τρία διακριτά επίπεδα τα οποία διαδοχικά εισάγουν θεμελιώδεις εκφράσεις: το επίπεδο των μεταδεδομένων (metadata layer), το επίπεδο του σχήματος (schema layer) και το επίπεδο της λογικής. Γλώσσες που υποστηρίζουν αυτή την αρχιτεκτονική φαίνονται στο επόμενο σχήμα:



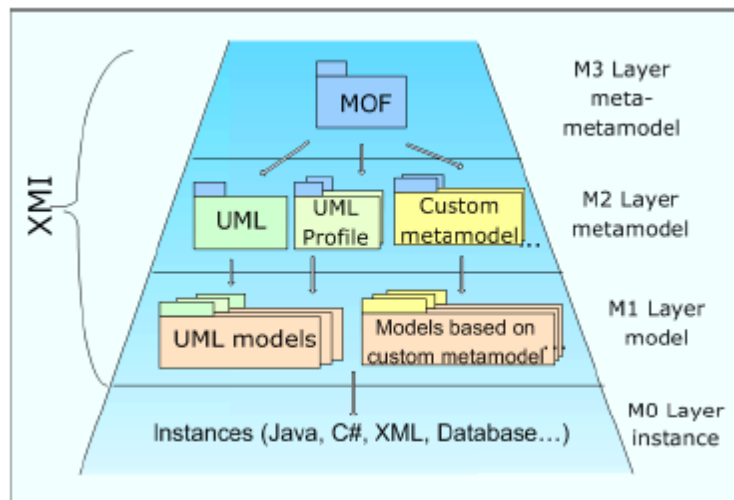
Σχήμα 3: Η αρχιτεκτονική του Σημασιολογικού Ιστού και η θέση της OWL

Στις σημερινές εφαρμογές, η διαλειτουργικότητα των κοινών δεδομένων επιτυγχάνεται με την χρήση της XML (Bray, Paoli, Sperberg-McQueen, & Maler, 2000). Όπως φαίνεται και στο προηγούμενο σχήμα η XML υποστηρίζει την σύνταξη ενώ η σημασιολογία παρέχεται από τις RDF, RDF Schema και κυρίως από την OWL. Την OWL θα την δούμε αναλυτικότερα στην επόμενη παράγραφο.

Η Αρχιτεκτονική Βασισμένη στο Μοντέλο (**M**odel **D**riven **A**rchitecture (OMG, MDA)) ορίζει τρία διαφορετικά επίπεδα αφαίρεσης από τα οποία μπορεί κανείς να δει το σύστημα. Για κάθε επίπεδο μπορεί να δοθεί μια αναπαράσταση ενός δοσμένου συστήματος. Με αυτόν τον τρόπο προκύπτουν τρία μοντέλα τα οποία είναι: Το

Μοντέλο Ανεξαρτήτως Υπολογισμών (Computation Independent Model), το Μοντέλο Ανεξαρτήτως Πλατφόρμας (Platform Independent Model) και το Μοντέλο Συγκεκριμένης Πλατφόρμας (Platform Specific Model). Στα τρία αυτά μοντέλα δεν υπάρχει λόγος να μπούμε σε περισσότερες λεπτομέρειες. Αυτό που θα δούμε είναι λίγο αναλυτικότερα την MDA αρχιτεκτονική στην οποία βασίστηκε και η περιγραφή των δεδομένων για το DBE.

Η MDA είναι βασισμένη σε μια αρχιτεκτονική μοντελοποίησης μεταδεδομένων τεσσάρων επιπέδων και σε πολλά συμπληρωματικά πρότυπα του OMG που φαίνονται στο παρακάτω σχήμα. Αυτά τα πρότυπα είναι το πρότυπο εκμετάλλευσης μετα-αντικειμένων MOF(Meta-Object Facility), η ενοποιημένη γλώσσα μοντελοποίησης UML (Unified Modeling Language) (Booch, Rumbaugh, & Jacobson, 1998) και το πρότυπο ανταλλαγής μεταδεδομένων XML, XMI (XML Metadata Interchange) (OMG, OMG XMI Specification, v1.2, 2002). Τα επίπεδα είναι: επίπεδο μετα-μέτα μοντέλου (M3), επίπεδο μεταμοντέλου (M2), επίπεδο μοντέλου (M1) και επίπεδο δεδομένων(στιγμιότυπων) (M0) .



Σχήμα 4: Η τεσσάρων επιπέδων MDA αρχιτεκτονική βασισμένη στο MOF.

Στην κορυφή της αρχιτεκτονικής είναι το μετα-μεταμοντέλο (MOF). Αυτό με τη σειρά του ορίζει μια αφηρημένη γλώσσα και ένα πλαίσιο για τον ορισμό, κατασκευή και διαχείριση μοντέλων τεχνολογικά ουδέτερων³³. Είναι η βάση για τον ορισμό

³³ Τεχνολογικά ουδέτερων σημαίνει πρακτικά ότι πρόκειται για μοντέλα που μπορούν να χρησιμοποιηθούν για την περιγραφή δεδομένων ανεξάρτητα από το σε ποιο σύστημα θα βρίσκονται αυτά, με τι πλατφόρμα ή τεχνολογίες θα σχετίζονται.

γλωσσών μονετελοποίησης όπως η UML ή ακόμα και το ίδιο το MOF. Το MOF επίσης ορίζει ένα πλαίσιο για την υλοποίηση αποθηκών για μεταδεδομένα δηλαδή μοντέλα. Για το MOF θα μιλήσουμε περισσότερο στο υπόλοιπο της παραγράφου. Ο βασικός σκοπός ύπαρξης των τεσσάρων επιπέδων με ένα κοινό μετα-μεταμοντέλο είναι η υποστήριξη πολλαπλών μεταμοντέλων και μοντέλων, η επεκτασιμότητά τους, η ενσωμάτωσή τους και η διαχείριση των μεταμοντέλων.

Πριν περάσουμε στο MOF, έγινε ξεκάθαρο με τα παραπάνω το τι ανάγκες ήρθε να καλύψει η Αρχιτεκτονική Βασισμένη στο Μοντέλο (MDA). Τα πρότυπα σε αυτήν την περιοχή είναι πλέον ώριμα και η χρήση τους είναι όχι μόνο αναγκαία αλλά και πολύτιμη για όσους έχουν σαν στόχο την διαλειτουργικότητα και την ενοποίηση των πληροφοριών είτε μιλάμε για τον ίδιο τον παγκόσμιο ιστό είτε για οποιοδήποτε άλλο δίκτυο με χρήστες και δεδομένα που να τα χαρακτηρίζει η φυσική ετερογένεια. Στη συνέχεια θα δούμε τους λόγους που οδήγησαν στην υιοθέτηση της συγκεκριμένης αρχιτεκτονικής και μεθοδολογίας. Θα πρέπει επίσης να τονίσουμε πως αυτό το πλαίσιο ακολουθήθηκε για την αποθήκευση και την διαχείριση των δεδομένων του συστήματος, για τον διαμοιρασμό τους μεταξύ των διαφορετικών κόμβων και με αυτόν τον τρόπο ήταν το ίδιο πλαίσιο που υπαγόρευσε την μεθοδολογία για την δημιουργία και επεξεργασία των ερωτήσεων του συστήματος. Χαρακτηριστικό είναι πως η γλώσσα ερωτήσεων του συστήματος η «Γλώσσα Ερώτησης Μοντέλων» QML (Query Modeling Language) είναι και αυτή βασισμένη στο MOF και την MDA αρχιτεκτονική. Περισσότερα όπως πάνω σε αυτό θα δούμε στο επόμενο κεφάλαιο το οποίο αναφέρεται αποκλειστικά στο σύστημα ερωτήσεων και τα διαφορετικά κομμάτια του.

Η Βάση Γνώσης του DBE εγκαταστάθηκε στους διαφορετικούς κόμβους μικρομεσαίων επιχειρήσεων που αποτελέσαν το ψηφιακό επιχειρηματικό οικοσύστημα. Ένα από τους βασικούς στόχους, όπως έχουμε ήδη δει, ήταν η υποδομή της Βάσης Γνώσης να μπορεί εγκαθίσταται σε οποιονδήποτε ενδιαφερόμενο κόμβο γεγονός που έκανε την ανεξαρτησία πλατφόρμας ζωτικής σημασίας. Κάθε κόμβος που συνεισφέρει με τους πόρους της στο δίκτυο, έχει και μια τοπική βασική υποδομή Βάσης Γνώσης την οποία ας ονομάσουμε Κόμβο Βάσης Γνώσης(KBG). Ο KBG βασίζεται σε έναν συνδυασμό ενός συστήματος διαχείρισης βάσης δεδομένων και ενός συστήματος αποθήκευσης MOF μεταδεδομένων. Όλοι μαζί οι KBG αποτελούν ένα δίκτυο ομότιμων κόμβων και συνεργάζονται για να

προσφέρουν έναν αποδοτικό τρόπο ανακάλυψης, περιγραφής, διαμοιρασμού, ανάλυσης και ενοποίησης της πληροφορίας, από όλες τις επιχειρήσεις. Η Βάση Γνώσης διαχειρίζεται διαφορετικά δεδομένα μέσα σε μοντέλα, περιγράφοντας την δομή και τις σχέσεις μεταξύ διαφορετικών πηγών πληροφορίας. Χρησιμοποιώντας μοντέλα για επιχειρήσεις και υπηρεσίες η Βάση Γνώσης επιτρέπει την αποτύπωση όλης της σημαντικής γνώσης. Τα μοντέλα δημιουργούνται και αποθηκεύονται στη Βάση Γνώσης ενσωματώνοντας τις λεπτομέρειες για την δομή και τις σχέσεις μεταξύ των οντοτήτων. Αυτές οι λεπτομέρειες είναι διαθέσιμες σε όλους τους χρήστες μέσα από μια σειρά εργαλείων που υλοποιήθηκαν στα πλαίσια του DBE, έξω όμως από τα πλαίσια της παρούσας εργασίας. Τα βασικά χαρακτηριστικά της υλοποίησης της Βάσης Γνώσης του DBE μετά την υιοθέτηση του παραπάνω πλαισίου είναι τα ακόλουθα:

- Βασισμένη σε πρότυπα: Υποστηρίζει τα πρότυπα του OMG για μοντελοποίηση και ανταλλαγή μοντέλων (πρότυπα για οργάνωση και ανταλλαγή της πληροφορίας, MOF 2.0 – XMI 1.4). Η υιοθέτηση υπαρχόντων προτύπων εξασφάλισε την συμμόρφωση με τις κατευθύνσεις της βιομηχανίας και επέτρεψε των διαμοιρασμό των μοντέλων μεταξύ διαφορετικών εργαλείων μοντελοποίησης.
- Επεκτασιμότητα. Η Βάση Γνώσης επεκτείνεται εύκολα για να καλύψει άλλα μοντέλα και πηγές πληροφορίας
- Υποστήριξη πολλαπλών μεταμοντέλων και μοντέλων. Η Βάση Γνώσης περιλαμβάνει πολλαπλά μεταμοντέλα για οντολογίες, επιχειρήσεις, υπηρεσίες και σύνθεση υπηρεσιών ενώ επιπλέον μεταμοντέλα μπορούν εύκολα να υλοποιηθούν και να υποστηριχθούν.
- Διαφορετικές όψεις. Υποστηρίζονται διαφορετικές όψεις της πληροφορίας δηλαδή διαφορετικά επίπεδα πρόσβασης στα ίδια μεταδεδομένα με διαφορετικές οπτικές.
- Ανεξαρτησία πλατφόρμας. Σχεδιασμένη και υλοποιημένη πλήρως για Java, η Βάση Γνώσης είναι ανεξάρτητη πλατφόρμας και έτσι μπορεί να εγκατασταθεί σε οποιαδήποτε επιχείρηση αδιαφορώντας για το οικείο λειτουργικό σύστημα.

Με τα προηγούμενα κάναμε σαφείς του λόγους για τους οποίους ακολουθήθηκε το πλαίσιο και η μεθοδολογία που περιγράψαμε στην αρχή της παραγράφου. Επίσης είναι σημαντικό ότι όλα αυτά παίζανε κυρίαρχο ρόλο στην κατασκευή του

υπερκείμενου δικτύου του DBE. Ήταν τα εργαλεία αλλά και ταυτόχρονα η μεθοδολογία τόσο για την κατασκευή όσο και για την εκμετάλλευση του υπερκείμενου δικτύου. Θυμίζουμε ότι μιλώντας για υπερκείμενο δίκτυο αναφερόμαστε στο δίκτυο που χτίζεται ανεξάρτητα από το φυσικό δίκτυο με βάση την αποθηκευμένη στους κόμβους γνώση και κυρίως την αντίστοιχη σημασιολογία. Επίσης λέγοντας εκμετάλλευση εννοούμε την δυνατότητα ανακάλυψης της γνώσης και ανάκτησής της κατ' απαίτηση και για συγκεκριμένο σκοπό. Στο τελευταίο περιλαμβάνεται όλο το σύστημα κατασκευής επεξεργασίας και προώθησης ερωτήσεων που θα δούμε στο επόμενο κεφάλαιο. Θα κλείσουμε την παρούσα παράγραφο με μια πιο αναλυτική αναφορά στο MOF. Έχουμε ήδη πει ότι σε αυτό βασίστηκε η κατασκευή όλων των μοντέλων του DBE καθώς και η γλώσσα ερωτήσεων του συστήματος.

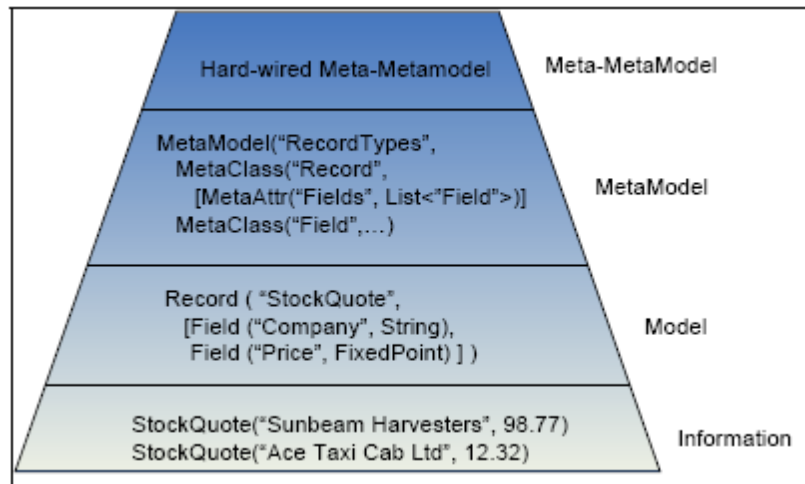
2.4.1. Το MOF

Με βάση όσα είπαμε στην αρχή της παραγράφου μιλώντας για την Αρχιτεκτονική Βασισμένη στο Μοντέλο (MDA) είναι εύκολο να παρακολουθήσουμε τον ορισμό του OMG σύμφωνα με το οποίο «το κεντρικό θέμα της προσέγγισης MOF για την διαχείριση μεταδεδομένων είναι η επεκτασιμότητα. Ο στόχος είναι να παρέχει ένα πλαίσιο που θα υποστηρίζει οποιουδήποτε τύπου μεταδεδομένα ενώ θα επιτρέπει να προστίθενται και νέοι τύποι όποτε απαιτείται. Για να το πετύχει αυτό, το ίδιο το MOF έχει μια αρχιτεκτονική επιπέδων μεταδεδομένων, βασισμένη στην κλασική τεσσάρων επιπέδων αρχιτεκτονική μοντελοποίησης που έγινε γνωστή σαν μέρος πρότυπων όπως τα ISO και CDIF. Το κλειδί τόσο για τις κλασικές όσο τη MOF αρχιτεκτονική είναι το επίπεδο μετα-μεταμοντελοποίησης το οποίο συνδέει μαζί το μοντέλο και το μεταμοντέλο». Τα τέσσερα παραδοσιακά επίπεδα όπως τα είδαμε κι νωρίτερα, είναι:

1. Το επίπεδο πληροφοριών. Αποτελείται από τα ίδια τα δεδομένα τα οποία θέλουμε να περιγράψουμε
2. Το επίπεδο μοντέλου. Περιέχει τα μεταδεδομένα που περιγράφουν τα δεδομένα του πρώτου επιπέδου. Αυτά είναι οργανωμένα ανεπίσημα σε μοντέλα.

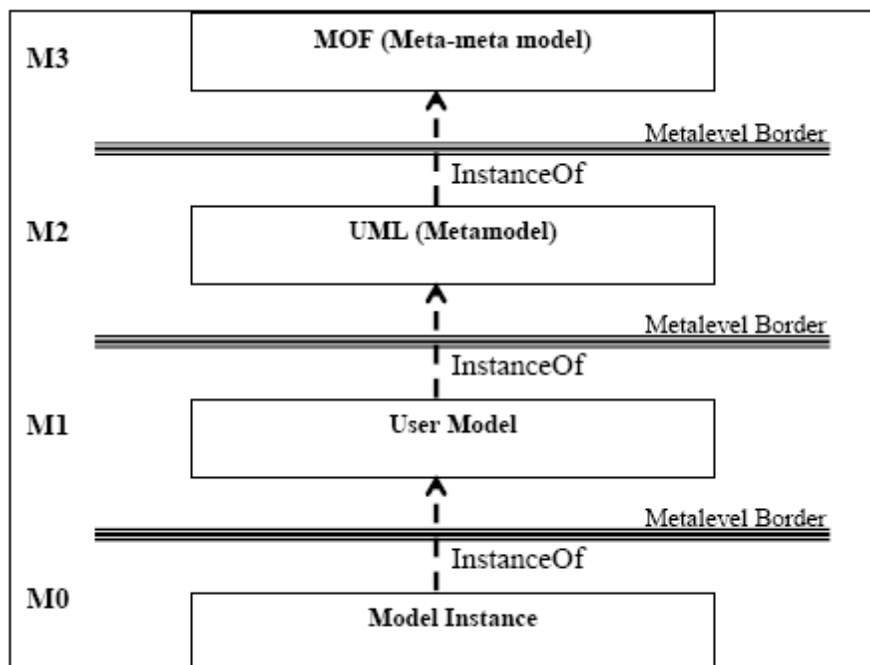
3. Το επίπεδο μεταμοντέλου. Περιέχει τις περιγραφές που ορίζουν τη δομή και τη σημασιολογία των μεταδεδομένων. Ένα μεταμοντέλο είναι ουσιαστικά μια αφηρημένη γλώσσα για την περιγραφή διαφορετικών τύπων δεδομένων.
4. Το επίπεδο μετα-μεταμοντέλου. Κατ' αντιστοιχία περιέχει τη δομή και τη σημασιολογία των μετα-μεταδεδομένων. Είναι με άλλα λόγια μια αφηρημένη γλώσσα για τον ορισμό διαφορετικού τύπου μεταδεδομένων.

Στο παράδειγμα του επόμενου σχήματος εξηγείται το πλαίσιο των τεσσάρων επιπέδων δείχνοντας την οργάνωση των μεταδεδομένων κάποιες απλές εγγραφές και μεταμοντέλο τύπων εγγραφών που χρησιμοποιούνται για να ορίσουν τη δομή και τη σημασιολογία των μεταδεδομένων.



Σχήμα 5: η αρχιτεκτονική τεσσάρων επιπέδων μεταδεδομένων

Ενώ στο παράδειγμα φαίνεται μόνο ένα μοντέλο και μεταμοντέλο ο κύριος σκοπός όπως είπαμε είναι η υποστήριξη πολλαπλών μοντέλων και μεταμοντέλων. Η προσέγγιση της μετα-μοντελοποίησης είναι ένα κοινός τρόπος οργάνωσης των μοντέλων που περιλαμβάνει περιγραφές σε επίπεδα το ένα πάνω στο άλλο. Οι έννοιες ενός επιπέδου έχουν τις αντίστοιχες περιγραφές τους στο επόμενο επίπεδο ή καλύτερα μετα-επίπεδο. Από την άλλη πλευρά θα μπορούσαμε να πούμε ότι ένα επίπεδο είναι ένα μοντέλο και το πιο κάτω επίπεδο είναι ένα στιγμιότυπο του μοντέλου. Σε σχέση με τον αντικειμενοστραφή προγραμματισμό το πιο χαμηλό επίπεδο περιέχει τα αντικείμενα ενός προγράμματος ενώ οι κλάσεις βρίσκονται στο πιο πάνω επίπεδο.



Σχήμα 6: Η MOF αρχιτεκτονική 4 επιπέδων

Η υποστηριζόμενη αρχιτεκτονική βασίζεται σε αυστηρή μεταμοντελοποίηση γεγονός που σημαίνει ότι όλα τα στοιχεία ενός επιπέδου έχουν προκύψει σαν στιγμιότυπα του πιο πάνω επιπέδου. Τόσο το MOF όσο και η UML προσφέρουν υποστήριξη για αντικειμενοστραφείς έννοιες και ο πυρήνας του MOF και της UML είναι δομικά ισοδύναμος. Τελικά, αφού το MOF χρησιμοποιείται για να ορίσει τον εαυτό του το επίπεδο πάνω από αυτό μπορεί επίσης να θεωρηθεί MOF καταλήγοντας σε ένα άπειρο αριθμό από επίπεδα MOF.

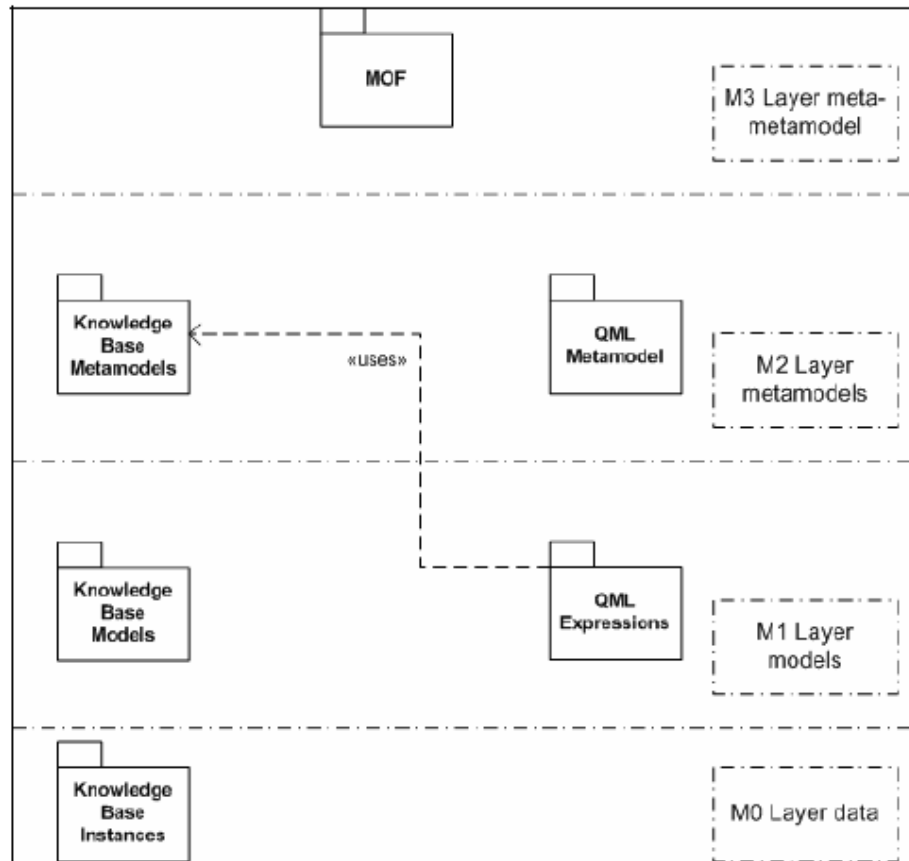
2.5. Η γλώσσα ερωτήσεων QML

Σύμφωνα με όσα έχουμε δει από την αρχή του κεφαλαίου τα MOF μοντέλα και τα στιγμιότυπά τους μπορούν να αντιστοιχηθούν σε XML έγγραφα (XMI³⁴) και ένας XML μηχανισμός ερωτήσεων μπορεί να ενσωματωθεί με την MOF τεχνολογία. Η XQuery και η XPATH είναι μερικές από τις πολλές γλώσσες ερωτήσεων για XML έγγραφα. Με αυτό τον τρόπο το πρόβλημα των ερωτήσεων στο MOF θα μπορούσε να αναχθεί σε ένα ήδη γνωστό πρόβλημα ερώτησης XML εγγράφων. Παρ' όλα αυτά η μη αντικειμενοστραφής φύση της XML περιορίζει την εκφραστική δύναμη μιας τέτοιας προσέγγισης ερωτήσεων. Η γλώσσα ερώτησης αντικειμένων (Object Query Language) είναι μια γλώσσα ερωτήσεων που βασίζεται στην SQL και έχει οριστεί από το OMG σαν τμήμα του προτύπου Δεδομένων Αντικειμένων (Atwood & Cattell, 1994). Το πρότυπο όμως δεν ορίζει την αφηρημένη σύνταξη της γλώσσας ή την επίσημη σημασιολογία της.

Η προσέγγιση που ακολουθήσαμε ήταν να ορίσουμε μια αντικειμενοστραφή γλώσσα πρόσβασης σε δεδομένα, την Γλώσσα Ερώτησης Μεταμοντέλων (**Query Metamodeling Language**) , χρησιμοποιώντας την ίδια μετα-γλώσσα που είχαμε χρησιμοποιήσει για τις γλώσσες περιγραφής της γνώσης στο DBE. Για να επιτύχουμε την μέγιστη συμβατότητα με τα υπάρχοντα πρότυπα βασιστήκαμε στην Γλώσσα Περιορισμού Αντικειμένων (Object Constraint Language) . Η QML ορίστηκε και υλοποιήθηκε από τον Γιώργο Κοτόπουλο (Kotopoulos, Kazasis, & Christodoulakis, 2007) στο ίδιο πλαίσιο που περιγράφουμε αλλά έξω από το αντικείμενο της παρούσας εργασίας. Για αυτό το λόγο, εδώ, θα αναφερθούμε μόνο στα γενικά χαρακτηριστικά της.

Η QML σύμφωνα με όσα έχουμε πει είναι ορισμένη σε M2 επίπεδο και σαν ένα MOF μεταμοντέλο. Επιτρέπει την δημιουργία ερωτηματικών εκφράσεων (M1 QML μοντέλα) χρησιμοποιώντας την πληροφορία των M2 μεταμοντέλων της Βάσης Γνώσης για την ανάκτηση M1 μοντέλων. Στο παρακάτω σχήμα φαίνεται η άμεση συσχέτιση των στοιχείων της QML με αυτά του MOF.

³⁴ Το XMI το είδαμε στο προηγούμενο κεφάλαιο και επιτρέπει την αντιστοίχιση αυτή πρακτικά και τεχνικά ώστε να μπορούμε να διαχειριζόμαστε οποιοδήποτε MOF έγγραφο σαν ένα XML έγγραφο.



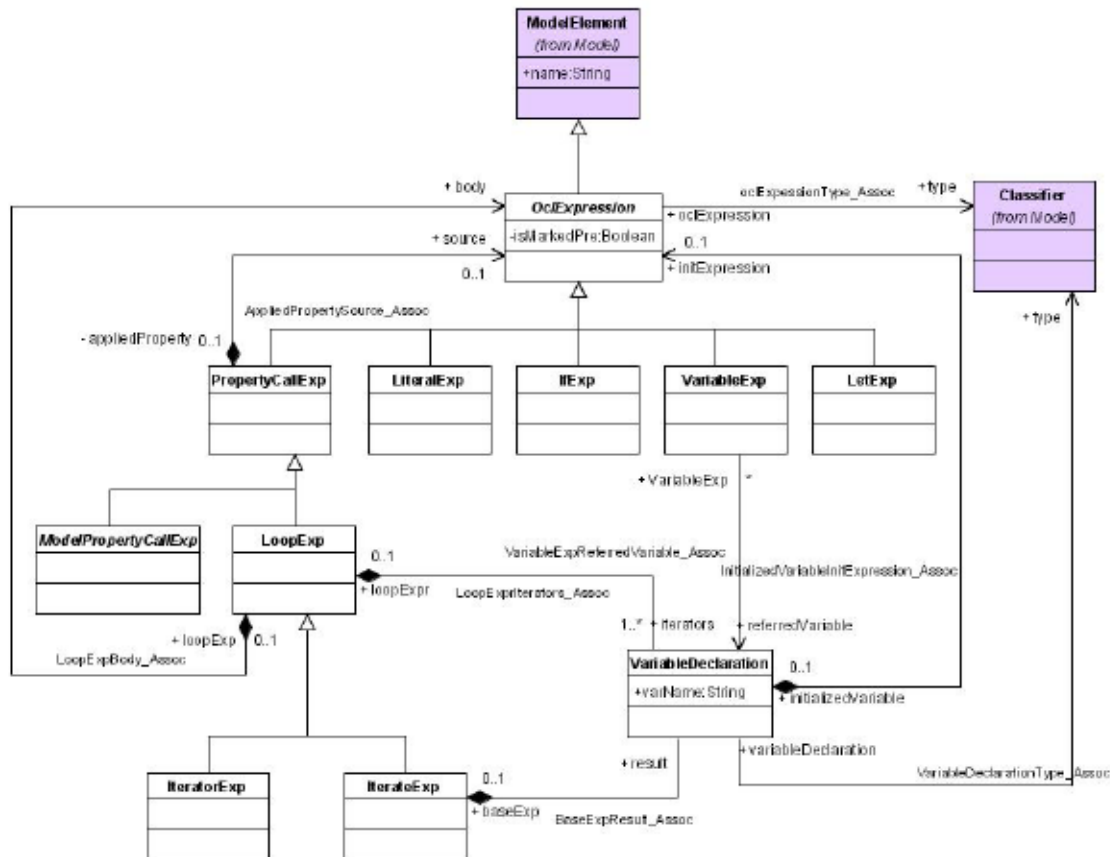
Σχήμα 7: Η διαδικασία πρόσβασης στα δεδομένα χρησιμοποιεί τη MOF αρχιτεκτονική. Η έγκυρες QML εκφράσεις είναι στιγμιότυπα του αντίστοιχου μεταμοντέλου. Αυτό επιτρέπει τη χρήση των στιγμιότυπων για όλα τα μοντέλα που υπάρχουν στη Βάση Γνώσης.

Ο πυρήνας του μεταμοντέλου της QML αποτελείται από δύο πακέτα: το πακέτο των εκφράσεων και το πακέτο των τύπων. Θα επικεντρώσουμε την περιγραφή μας στο πακέτο των εκφράσεων συνοψίζοντας τις δυνατότητες που παρέχει η QML για την δημιουργία ερωτήσεων ενώ τα υπόλοιπα στοιχεία περιλαμβάνονται στο παράρτημα [1].

Πακέτο Εκφράσεων

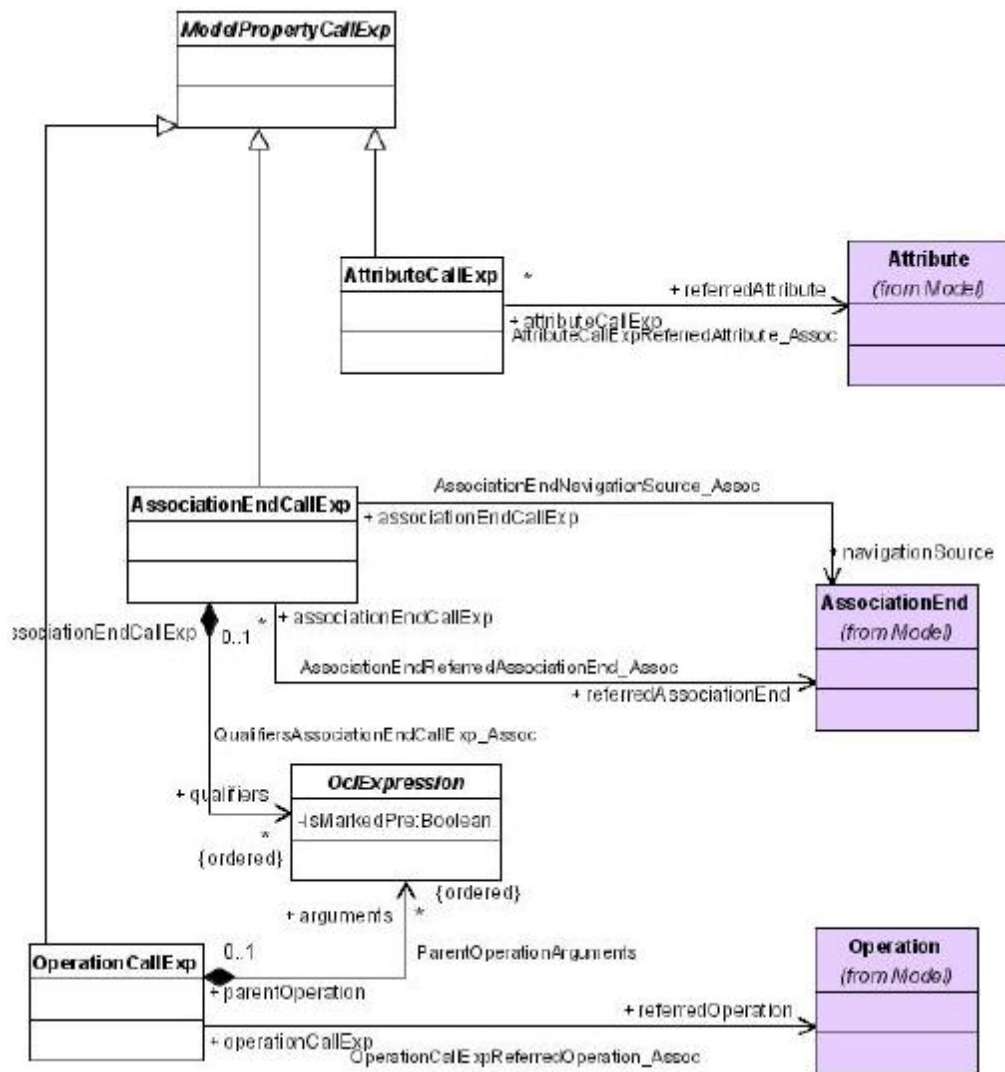
Στο επόμενο σχήμα υπάρχει ο πυρήνας του πακέτου των Εκφράσεων. Η βασική δομή του πακέτου αποτελείται από τις κλάσεις `OclExpression`, `PropertyCallExpr` and `VariableExpr`. Ένα `OclExpression` έχει πάντα τύπο ο οποίος συνήθως προκύπτει από κάποιον έτοιμο. Κάθε `PropertyCallExpr` έχει ακριβώς μια πηγή που αναγνωρίζεται από ένα `OclExpression`. Προκειμένου να μπορούμε να εκφράσουμε περιορισμούς ενσωματωμένους στο μοντέλο, πάνω σε κάποιο στοιχείο του μοντέλου, η δομή του

MOF μας επιβάλλει να εξειδικεύσουμε μια από τις ModelElement κλάσης του MOF μέσω ενός OclExpression. Σε αυτό το σημείο χρησιμοποιούμε τον όρο ιδιότητα “property” που είναι μια γενίκευση του χαρακτηριστικού “Feature”, “AssociationEnd” και έχει οριστεί για να διατρέχει OCL συλλογές από πράξεις.



Σχήμα 8: Το πακέτο εκφράσεων της QML

Ένα `ModelPropertyCallExpr` γενικεύει όλες τις κλήσεις ιδιοτήτων που αναφέρονται σε “Features” ή “AssociationEnds”. Στο επόμενο σχήμα ορίζονται όλοι οι διαφορετικοί `ModelPropertyCallExpr` τύποι. Στο υπόλοιπο του πακέτου υπάρχουν όλες οι διαφορετικές υποκλάσεις του `PropertyCallExpr`. Επίσης μπορεί κανείς να καταλάβει ότι μια QML έκφραση ξεκινάει πάντα είτε με μια μεταβλητή είτε με ένα συγκεκριμένο όρο, πάνω στον οποίο εφαρμόζεται αναδρομικά μια ιδιότητα.



Σχήμα 9: Το ModelPropertyCallExp στο Πακέτο Εκφράσεων

IterateExp

Πρόκειται για μια έκφραση που εφαρμόζει τον υπολογισμό που περιγράφεται στο κύριο μέρος της για κάθε στοιχείο της συλλογής. Λειτουργεί επαναληπτικά πάνω στη συλλογή και υπολογίζει το αποτέλεσμα της έκφρασης.

IteratorExp

Είναι μια έκφραση που αντιπροσωπεύει όλες τις ορισμένες στη συλλογή λειτουργίες που χρησιμοποιούν επανάληψη.

LoopExp

Είναι μια έκφραση που αντιπροσωπεύει τη δομή ενός βρόγχου επανάληψης πάνω σε μια συλλογή. Χρησιμοποιεί έναν επαναλήπτη και υπολογίζει την δοσμένη έκφραση για κάθε ένα στοιχείο της συλλογής.

OclExpression

Είναι μια έκφραση που μπορεί να υπολογιστεί σε ένα δοσμένο περιβάλλον. Το OclExpression είναι μια αφηρημένη υπερ-κλάση όλων των άλλων εκφράσεων στο μεταμοντέλο. Είναι το πιο πάνω στοιχείο στο πακέτο των εκφράσεων. Κάθε OclExpression έχει έναν τύπο που μπορεί καθοριστεί με ανάλυση της έκφρασης και των συμφραζομένων της. Ο υπολογισμός μιας έκφρασης έχει σαν αποτέλεσμα κάποια τιμή. Επίσης το περιβάλλον μιας έκφρασης ορίζει ποια στοιχεία του μοντέλου είναι ορατά και μπορούν να αναφερθούν μέσα στην έκφραση.

PropertyCallExp

Ένα PropertyCallExp είναι μια έκφραση που αναφέρεται σε μια ιδιότητα και έχει κάποια πηγή. Η τιμή του αποτελέσματός της είναι ο υπολογισμός της αντίστοιχης ιδιότητας. Το PropertyCallExp είναι μια αφηρημένη μετα-κλάση και η πηγή της είναι το στιγμιότυπο που πραγματοποιεί την κλήση στην ιδιότητα.

VariableDeclaration

Το VariableDeclaration δηλώνει μια μεταβλητή και τη συνδέει με κάποιον τύπο. Η μεταβλητή μπορεί να χρησιμοποιηθεί σε εκφράσεις που βρίσκονται στην εμβέλειά της. Η μετα-κλάση VariableDeclaration αντιπροσωπεύει τις μεταβλητές τις έκφρασης και του αποτελέσματος και τις μεταβλητές που ορίζονται χρησιμοποιώντας μια Let έκφραση.

VariableExp

Η VariableExp είναι μια έκφραση που αποτελείται από την αναφορά σε κάποια μεταβλητή σαν αυτές που περιγράψαμε προηγουμένως.

AssociationEndCallExp

Μια AssociationEndCallExp είναι μια αναφορά σε ένα AssociationEnd ορισμένο σε ένα MOF model. Χρησιμοποιείται για να ορίσει αντικείμενα συνδεδεμένα με ένα αντικείμενο στόχο από μια συσχέτιση. Η έκφραση αναφέρεται σε αυτά τα αντικείμενα στόχους με το όνομα του ρόλου του άκρου της συσχέτισης που βρίσκεται συνδεδεμένη η κλάση στόχος.

AttributeCallExp

Ένα AttributeCallExpression είναι μια αναφορά σε ένα Attribute ορισμένου στο MOF μοντέλο και υπολογίζει την τιμή ενός attribute.

OperationCallExp

Ένα OperationCallExp αναφέρεται σε μια λειτουργία-πράξη. Η έκφραση μπορεί να περιέχει μια λίστα από εκφράσεις σαν ορίσματα αν η λειτουργία έχει παραμέτρους. Σε αυτή τη περίπτωση ο αριθμός και οι τύποι των παραμέτρων πρέπει να ταιριάζουν με τα ορίσματα.

IfExp

Ένα IfExp οδηγεί σε μία από τις δύο εναλλακτικές εκφράσεις με βάση την υπολογιζόμενη τιμή της συνθήκης.

LetExp

Ένα LetExp είναι μια ειδική έκφραση που ορίζει μια νέα μεταβλητή με κάποια αρχική τιμή. Μια μεταβλητή που ορίζεται από ένα LetExp δεν μπορεί να αλλάξει την τιμή της. Η μεταβλητή είναι ορατή μέσα στην έκφραση.

LiteralExp

Ένα LiteralExp είναι μια έκφραση χωρίς ορίσματα που να παράγουν μια τιμή. Γενικά η τιμή του αποτελέσματος είναι ταυτόσημη με το σύμβολο της έκφρασης. Δεν θα μπορούμε σε περισσότερες λεπτομέρειες σε αυτό το σημείο για τον ορισμό της QML γιατί είναι έξω από τα όρια της παρούσας εργασίας. Όσα παρουσιάσαμε θα είναι πολύ πιο κατανοητά στον αναγνώστη που είναι εξοικειωμένος με την OCL. Παρ' όλα αυτά θα κλείσουμε την παράγραφο με κάποια παραδείγματα από QML

εκφράσεις τα οποία είναι σίγουρα αρκετά πιο καταληπτά και συγκρίσιμα είτε με OCL εκφράσεις είτε και με XQuery ερωτήσεις.

2.5.1. QML Εκφράσεις και σημασιολογία

Σε αυτό το σημείο θα δούμε κάποια παραδείγματα ερωτήσεων γραμμένα με τη μορφή QML εκφράσεων. Οι ερωτήσεις που θα ακολουθήσουν βασίζονται στο Μεταμοντέλο Σημασιολογικών Υπηρεσιών (SSL) το οποίο, όπως έχουμε δει, προορίζεται για την σημασιολογική περιγραφή υπηρεσιών. Τα βασικά στοιχεία του SSL που θα δούμε είναι τα εξής:

- **SemanticPackage**: αναφέρεται στο σύνολο των περιγραφών για μια υπηρεσία
- **ServiceProfile**: είναι το μοντέλο σύμφωνα με το οποίο θα περιγραφεί σημασιολογικά η υπηρεσία
- **Attribute**: αφορά μια τιμή σημασιολογικής πληροφορίας για κάποιο συγκεκριμένο μοντέλο
- **Functionality**: περιγράφει τι μπορεί να κάνει ο χρήστης μιας υπηρεσίας αλληλεπιδρώντας με τον παροχέα της υπηρεσίας.

Μια απλή QML έκφραση

Στη συνέχεια βλέπουμε τη δήλωση μιας QML έκφρασης. Ας θεωρήσουμε λοιπόν την ακόλουθη δήλωση:

```
Context SSL::ServiceProfile simpleQuery : SemanticPackage
query:
Attribute.name = "HotelName"
```

Η δήλωση ορίζει μια ερώτηση με όνομα "simpleQuery". Η ερώτηση τίθεται στο μεταμοντέλο της SSL το οποίο είναι ένα από τα μοντέλα που υποστηρίζεται από τη Βάση Γνώσης. Η ερώτηση ανακτά όλες τις υπηρεσίες που έχουν ένα χαρακτηριστικό με όνομα "HotelName". Η ερώτηση αναφέρεται στην ακολουθία "SSL", "ServiceProfile" και επίσης ορίζεται και ο τύπος του αποτελέσματος της ερώτησης ο οποίος είναι ένα "SemanticPackage". Στο σώμα της ερώτησης, βέπουμε ότι θα πρέπει να ταιριάζει σε στιγμιότυπο του ServiceProfile και η έκφραση προς υπολογισμό είναι η λειτουργία του '='. Η λειτουργία αυτή βρίσκεται ορισμένη στη βασική βιβλιοθήκη του OCL σαν μια λειτουργία του βασικού τύπου String. Αυτή η

λειτουργία έχει δύο ορίσματα. Το πρώτο βοηθάει στον εντοπισμό της θέσης μέσω της κλάσης "ServiceProfile". Το δεύτερο είναι απλά μια τιμή. Πιο αναλυτικά, για τον εντοπισμό του περιβάλλοντος της ερώτησης, μέσω του "ServiceProfile" το "Attribute" αποτελεί ένα AssociationEnd της συσχέτισης μεταξύ των κλάσεων ServiceProfile και ServiceAttribute. Τελικά το όνομα εμφανίζεται σαν χαρακτηριστικό της κλάσης ServiceAttribute.

Δεν θα μπορούμε σε περισσότερες λεπτομέρειες διότι είναι έξω από τους στόχους του παρόντος κεφαλαίου αλλά και του αντικειμένου μας. Θα κλείσουμε όμως με την απλή παράθεση μιας έκφρασης ερώτησης λίγο πιο περίπλοκης ώστε να γίνει πιο εμφανής η δομή μιας QML ερώτησης αλλά και η εκφραστική της δυνατότητα.

```
Context SSL:ServiceProfile complexQuery: SemanticPackage
query:
Functionality->select (name="CreditCardPayment") ->
exist (input.name="CreditCardNumber")
and
Attribute->select (name="Address") ->
exists (type="ODM::HotelDomain::Address" and getTypeClassInstance() ->
select (type.name="City") ->exists (TheDTPRange.lexicalForm="Chania"))
```

Η παραπάνω ερώτηση αναφέρεται πάλι σε ένα μεταμοντέλο SSL και ανακτά όλες τις υπηρεσίες της περιοχής των ξενοδοχείων που βρίσκονται στα Χανιά και παρέχουν λειτουργίες πληρωμής μέσω πιστωτικής κάρτας.

Κεφάλαιο 3 – Ανάπτυξη Συστήματος και Συνεισφορά

3.1. Εισαγωγή

Στο κεφάλαιο που ακολουθεί θα δούμε αναλυτικότερα τη χρήση του τεχνικού πλαισίου που περιγράψαμε στο προηγούμενο κεφάλαιο, προκειμένου να παρουσιάσουμε τα διάφορα σημεία συνεισφοράς αυτής της εργασίας. Είδαμε πως κύριος άξονας ήταν η ανάπτυξη ενός ομότιμου δικτύου από ετερογενείς κόμβους. Βασικό στοιχείο η περιγραφή και διαχείριση της γνώσης με ενιαίο τρόπο. Όπως είδαμε και στην εισαγωγή μιλάμε για γνώση και όχι για δεδομένα γιατί με τα σημερινά εργαλεία τα δεδομένα έχουν αποκτήσει την ικανότητα να φέρουν μαζί τους τη λογική και τη σημασιολογία τους. Το σημαντικότερο εργαλείο είναι φυσικά οι οντολογίες και το αποφασιστικό βήμα για την υποστήριξή τους είναι η γλώσσα περιγραφής τους που στην περίπτωση μας ορίστηκε σαν μέρος της παρούσας εργασίας σαν ένα μετά-μοντέλο στο πλαίσιο που περιγράψαμε στο προηγούμενο κεφάλαιο (§ 2.2.2), το ODM. Την αναλυτική περιγραφή του ODM θα δούμε στη συνέχεια του κεφαλαίου (§ 3.3). Με αυτόν τον τρόπο χτίστηκε μια Βάση Γνώσης με τη δυνατότητα παροχής διαλειτουργικότητας μεταξύ των διαφορετικών χρηστών του συστήματος. Όπως είδαμε και στην αρχή το σύστημα είναι βασισμένο στις δημοσιοποιημένες υπηρεσίες οι οποίες εξυπηρετούν την ανάγκη της παροχής πραγματικών υπηρεσιών από επιχειρήσεις προς πελάτες και της εκμετάλλευσης της παρεχόμενης λειτουργικότητας. Στη συνέχεια του κεφαλαίου θα δούμε πως η συνεισφορά μας εντοπίζεται:

- στην ανάπτυξη του ομότιμου δικτύου
- στον ορισμό της γλώσσας περιγραφής των δεδομένων του συστήματος
- στην κατασκευή και διαχείριση της Βάσης Γνώσης
- στην εκμετάλλευση των παραπάνω για την παροχή υπηρεσιών
- στη δημιουργία μιας υψηλού επιπέδου μηχανής εκτέλεσης των ερωτήσεων
- στη μελέτη και βελτίωση της διαδικασίας επεξεργασίας και προώθησης των ερωτήσεων

Το τελευταίο σημείο εκτός από το καθαρά τεχνικό κομμάτι που αφορά την ίδια την επεξεργασία των ερωτήσεων, αναφέρεται στη διαδικασία ανακάλυψης τόσο των παραπάνω υπηρεσιών όσο και των αντίστοιχων κόμβων. Το πρόβλημα φυσικά

μπορεί να παρουσιασθεί και αντίστροφα σαν αυτό της παροχής προτάσεων προς τους πελάτες για επιθυμητές σε αυτούς υπηρεσίες και συνεργάτες. Έχοντας μια Βάση Γνώσης με τα χαρακτηριστικά που αναλυτικά περιγράψαμε, η πρόκληση που προκύπτει είναι αυτή της κάλυψης της ανάγκης για αναζήτηση και ανάκτηση γνώσης. Συνεχίζοντας να υποστηρίζουμε την διαλειτουργικότητα που παρέχει η αρχιτεκτονική και η υποδομή του συστήματος έπρεπε να αναπτύξουμε ένα ολοκληρωμένο σύστημα δημιουργίας και επεξεργασίας ερωτήσεων που να υπηρετεί τις ίδιες ανάγκες. Για το σκοπό αυτό παρουσιάσαμε (§ 2.5) την γλώσσα του συστήματος (Γλώσσα Ερώτησης Μεταμοντέλων - QML) ενώ τώρα θα δούμε τους μηχανισμούς επεξεργασίας των ερωτήσεων. Η όλη ανάπτυξη ήταν ανεξάρτητη πλατφόρμας και τελικού συστήματος αποθήκευσης των δεδομένων. Το κεφάλαιο κλείνει επεκτείνοντας την ανάγκη διαχείρισης των ερωτήσεων σε επίπεδο δικτύου μέσω του μηχανισμού προώθησης των ερωτήσεων. Στο τέλος του κεφαλαίου θα επικεντρωθούμε στο πρόβλημα αυτό και στη μεθοδολογία που ακολουθήθηκε ενώ θα κάνουμε και μια εισαγωγή στο θέμα της επεξεργασίας των ερωτήσεων για την εκτίμηση του κόστους τους, που θα αποτελέσει το αντικείμενο του επόμενου 4^{ου} κεφαλαίου.

3.2. Ανάπτυξη Ομότιμου δικτύου

Όπως είδαμε ήδη από την παράγραφο 2.2.1 στόχος μας δεν ήταν η υλοποίηση του φυσικού δικτύου εξ αρχής και για αυτό το λόγο βασιστήκαμε στην πλατφόρμα FADA. Η πλατφόρμα παρέχει τα βασικά χαρακτηριστικά για την ανάπτυξη ενός φυσικού δικτύου ενώ είναι ανοικτού κώδικα ώστε να επιδέχεται προσαρμογές στις εκάστοτε ανάγκες. Με τον τρόπο αυτό έγιναν παρεμβάσεις και προσθήκες προκειμένου να δημιουργηθεί μια πλατφόρμα που θα επέτρεπε το στήσιμο ενός φυσικού δικτύου με τα χαρακτηριστικά που θέλαμε (§ 2.3.1). Έχοντας σαν βάση το τεχνικό πλαίσιο που είδαμε στο προηγούμενο κεφάλαιο, οι στόχοι που επιτεύχθηκαν ήταν οι ακόλουθοι:

- ❖ Η εξοικείωση με τις διάφορες κατηγορίες δικτύων και η χρήση σε χαμηλό επίπεδο ενός ομότιμου δικτύου. Είναι σημαντικό το ότι ήρθαμε σε επαφή με την υποδομή του δικτύου στη βάση της κατασκευής του γιατί με αυτόν τον τα τρόπο υπήρχε η δυνατότητα παρέμβασης στα ίδια τα χαρακτηριστικά του δικτύου ενώ ήταν πιο διάφανη η όλη λειτουργία του

- ❖ Η εγκατάσταση του δικτύου σε πραγματικό περιβάλλον. Αυτό έγινε τόσο στο εργαστήριο Διανεμημένων Πληροφοριακών Συστημάτων και Εφαρμογών όπου υπήρχε το δίκτυο εγκατεστημένο καθ' όλη τη διάρκεια του προγράμματος όσο και εκτός εργαστηρίου στα μηχανήματα διαφόρων συνεργατών του προγράμματος σε διάφορα σημεία του εξωτερικού. Επίσης δοκιμή σε πραγματικές συνθήκες έγινε κατά τη διάρκεια των αξιολογήσεων του προγράμματος με τη συμμετοχή παραπάνω και εξειδικευμένων χρηστών
- ❖ Η υλοποίηση και δημοσίευση υπηρεσιών χρησιμοποιώντας την μεθοδολογία που περιγράφηκε κατά την παρουσίαση του FADA. Έγινε υλοποίηση συγκεκριμένων υπηρεσιών που ενσωματώθηκαν στο λογισμικό των κόμβων του συστήματος. Η υλοποίηση έγινε σε Java και οι υπηρεσίες ικανοποιούσαν απαιτήσεις της παρούσας εργασίας που θα αναλυθούν παρακάτω. Το σημαντικό στοιχείο όμως ήταν η ίδια η υλοποίηση των υπηρεσιών η οποία έγινε τόσο ακολουθώντας την φιλοσοφία των εξουσιοδοτών όσο και την τεχνολογία των υπηρεσιών πάνω από το διαδίκτυο (web services). Γενικά το πλαίσιο που ακολουθήθηκε για την υποστήριξη της λειτουργικότητας ήταν προσανατολισμένο στις υπηρεσίες και όλη η υλοποίηση ενσωματώθηκε σε υπηρεσίες προκειμένου να δημοσιευθεί και να είναι διαθέσιμη σε όλους τους πελάτες του δικτύου.

Οι υπηρεσίες που υλοποιήθηκαν αφορούσαν την επικοινωνία των κόμβων μεταξύ τους και την τήρηση στοιχείων του φυσικού δικτύου που ήταν χρήσιμα στη συνέχεια στην προώθηση ερωτήσεων και τη συλλογή απαντήσεων και αποτελεσμάτων. Αυτό ήταν από τους βασικούς μας στόχους καθώς σε αυτήν τη διαδικασία βασίστηκε όλη η διαχείριση και ανάκτηση της γνώσης για τις ανάγκες του συστήματος. Σημαντικό είναι πως όλη τελική η υποδομή του δικτύου προσέφερε εύκολη προώθηση των ερωτήσεων με την μέθοδο της κατάκλισης. Εκεί βασίστηκε η τροποποίηση στη συνέχεια της υλοποίησης προκειμένου να ενσωματωθούν πιο «έξυπνες» τεχνικές και να γίνει εκμετάλλευση ολόκληρου του αρχικού θεωρητικού πλαισίου(σημασιολογικά δεδομένα, πολλαπλές οντολογίες κ.ά.).

Εδώ είναι το κατάλληλο σημείο να σημειώσουμε ότι το να μην είναι η τοπολογία του δικτύου γνωστή εκ των προτέρων οδηγεί φυσικά σε κάποια μείωση της δυναμικής απόδοσης ενός τέτοιου συστήματος. Παρ' όλα αυτά θεωρήθηκε

σημαντικότερη η ευελιξία και ή αξιοπιστία του συστήματος. Το σημαντικότερο θέμα που προκύπτει είναι φυσικά η προώθηση των αιτημάτων αναζήτησης, στα οποία στο εξής θα αναφερόμαστε απλά ως ερωτήσεις. Πολλές προτάσεις έχουν γίνει για την καλύτερη αντιμετώπιση του συγκεκριμένου θέματος με πρώτη απ' όλες αυτή της πλήρους κατάκλισης³⁵ του δικτύου. Η διαδικασία της κατάκλισης, ενώ αποτελεί μια πρακτική που έχει αποτελέσματα, στην πράξη προβάλλει το θέμα της διάδοσης και απάντησης ερωτήσεων σαν μια πρόκληση η οποία δεν έχει εύκολη απάντηση όταν το ζητούμενο είναι η καλύτερη απόδοση. Τα ιδιαίτερα χαρακτηριστικά κάθε δικτύου παίζουν καθοριστικό ρόλο ενώ άλλες παράμετροι όπως ο τρόπος περιγραφής των δεδομένων του συστήματος αλλά και η διαδικασία επεξεργασίας των ερωτήσεων είναι εξ ίσου σημαντικές. Όλα αυτά αποτέλεσαν και την αφορμή για την παρούσα εργασία. Όπως αναφέρθηκε και στο πρώτο κεφάλαιο η επεξεργασία των ερωτήσεων καθώς και η προώθησή τους ήταν ο βασικός μας στόχος.

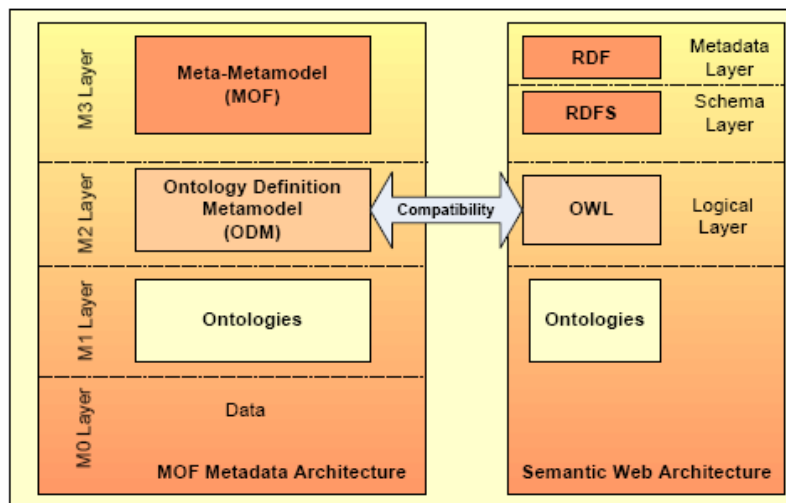
3.3. Το ODM σαν γλώσσα περιγραφής οντολογιών

Σε αυτό το σημείο θα περιγράψουμε το μεταμοντέλο ορισμού οντολογιών (ODM). Θα δούμε παραδείγματα της χρήσης των διαφόρων εννοιών ενώ ο βασικός στόχος ήταν η χρήση του MOF για τον ορισμό ενός μεταμοντέλου συμβατού με την OWL. Για το λόγο αυτό όχι απλώς υπάρχουν συνεχής αναφορές στις έννοιες τις OWL αλλά δεδομένου ότι όλες οι τελικές δυνατότητες του μεταμοντέλου είναι πλήρως συμβατές με αυτές τις OWL, για τις ανάγκες της παρούσας εργασίας θα μπορούσε να γίνει αντικατάσταση του ODM με την OWL για τους χρήστες που είναι εξοικειωμένοι με αυτήν. Η παρουσίαση του ODM γίνεται ώστε να υπάρχει συνέχεια στην παρουσίαση του πλαισίου της παρούσας εργασίας όπως αυτή γίνεται από τις προηγούμενες παραγράφους. Για τα επόμενα όμως κεφάλαια ο αναγνώστης μπορεί με σιγουριά να χρησιμοποιεί τις έννοιες τις OWL χωρίς κίνδυνο να υπάρχει κάποια αναντιστοιχία.

3.3.1. Μετα-μοντελοποίηση και ODM

³⁵ Αναφερόμαστε στην διαδικασία του «κατακλυσμού» του δικτύου με κάθε ερώτηση που δημιουργείται σε έναν οποιονδήποτε κόμβο, γνωστή ως flooding.

Η αρχιτεκτονική τεσσάρων επιπέδων που είδαμε μέχρι τώρα είναι διαφορετική από την αρχιτεκτονική μεταδεδομένων του σημασιολογικού ιστού η οποία όχι μόνο δεν είναι σταθερή αλλά ορίζει και τα δικά της επίπεδα με διαφορετική σημασιολογία από αυτήν του MOF. Αυτές οι διαφορές επέβαλλαν να ορίσουμε το δικό μας μεταμοντέλο περιγραφής μεταδεδομένων. Επίσης καθώς το MOF δεν παρέχει την εκφραστική δύναμη του RDFS έπρεπε να δοθεί ιδιαίτερη προσοχή ώστε να επιτευχθεί η απαιτούμενη εκφραστικότητα στο ODM. Σημαντικό είναι ότι το ODM ανταποκρίνεται σε όλους τους αντικειμενικούς στόχους που είχε το OMG στη συγκεκριμένη περιοχή.



Σχήμα 10: Η προσέγγιση μοντελοποίησης του ODM

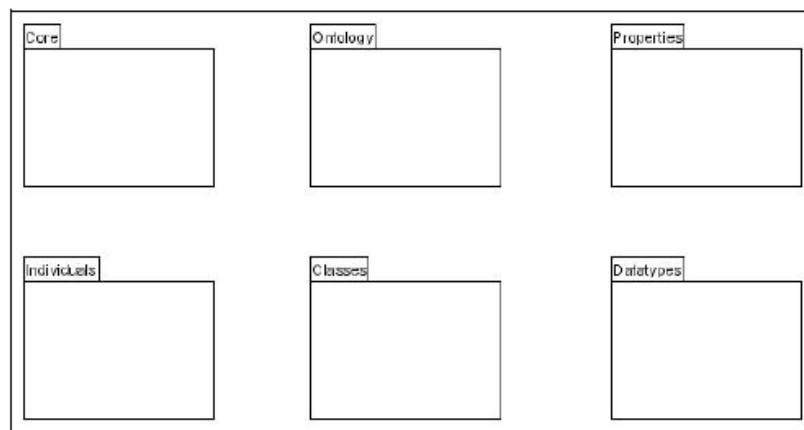
Το παραπάνω σχήμα δείχνει το ODM σε σχέση με την MOF αρχιτεκτονική μεταδεδομένων αλλά και που στέκεται στην αρχιτεκτονική μεταδεδομένων του σημασιολογικού ιστού. Το Μεταμοντέλο Ορισμού Οντολογιών βρίσκεται στο M2 επίπεδο μοντέλων και τα στιγμιότυπά του (M1 επιπέδου μοντέλα) είναι οι οντολογίες τις οποίες μπορούν να ορίσουν οι ειδικοί και οι χρήστες. Θα πρέπει να σημειωθεί ότι σύμφωνα με τον ορισμό του ODM μπορεί κανείς να περιγράψει τόσο στιγμιότυπα όσο και κλάσεις στο M1 επίπεδο. Αυτό έγινε για δύο λόγους:

- για να είναι να είναι κοινός ο μηχανισμός αναπαράστασης των στιγμιότυπων και των κλάσεων και
- για να μπορούν να χρησιμοποιηθούν υπάρχουσες οντολογίες που είναι περιγεγραμμένες σε OWL(αφού η OWL χρησιμοποιεί τον ίδιο μηχανισμό για να περιγράψει κλάσεις και στιγμιότυπα μέσα στην ίδια οντολογία).

Για τον ορισμό του ODM χρησιμοποιήθηκαν τα ακόλουθα βασικά στοιχεία του MOF:

1. MOF **Κλάση (Class)** για την μοντελοποίηση των μετα-αντικειμένων
2. MOF **Χαρακτηριστικό (Attribute)** το οποίο μοντελοποιεί την νοητή υποδοχή τιμής ενός μετα-αντικειμένου
3. MOF **Συσχέτιση (Association)** η οποία μοντελοποιεί τις δυαδικές συσχετίσεις μεταξύ μετα-αντικειμένων
4. MOF **Τύπος (Datatype)** δεδομένων ο οποίος μοντελοποιεί άλλα δεδομένα (όπως βασικούς ή εξωτερικούς τύπους)
5. MOF **Πακέτο (Package)** για την οργάνωση των μοντέλων

Με βάση τα παραπάνω η οργάνωση των μοντέλων στο ODM γίνεται στα εξής πακέτα:



Σχήμα 11: Προβολή των πακέτων του ODM

Δεν θα μπούμε σε περισσότερες λεπτομέρειες για τα πακέτα ή τις αφηρημένες έννοιες του ODM μιας και αυτές επισυνάπτονται στο παράρτημα Α ενώ δεν επηρεάζουν άμεσα τα όσα θα αναφερθούν στο υπόλοιπο της εργασίας. Το ίδιο ισχύει και για τον γενικό ορισμό μιας οντολογίας σε πιο πάνω επίπεδο. Στο υπόλοιπο της παραγράφου ασχολούμαστε με τα εσωτερικά δομικά στοιχεία μιας οντολογίας και τα πακέτα που μ ενδιαφέρουν για τη συνέχεια.

Κλάσεις

Η έννοια της κλάσης είναι θεμελιώδης για το ODM. Μια κλάση αναπαριστά μια ομάδα από όμοια αντικείμενα μέσα σε μια οντολογία. Με αυτόν τον τρόπο μια

κλάση συσχετίζεται με ένα σύνολο από άτομα³⁶ το οποίο και καλείται προέκταση της κλάσης. Τα άτομα τα οποία ανήκουν στην προέκταση μια κλάσης ονομάζονται τα στιγμιότυπα της κλάσης συσχετιζόμενα με την προέκταση της κλάσης. Μια κλάση έχει κάποιο συγκεκριμένο νόημα το οποίο σχετίζεται αλλά δεν ταυτίζεται με την προέκτασή της. Για αυτόν το λόγο δύο κλάσης μπορεί να εμφανιστούν να έχουν την ίδια προέκταση αλλά παρ' όλα αυτά να παραμένουν δύο διαφορετικές κλάσεις.

Ορισμός Κλάσεων

Μια ODM κλάση μπορεί να οριστεί είτε με το όνομα μια κλάσης είτε καθορίζοντας την προέκταση μια ανώνυμης κλάσης. Όπως και η OWL το ODM αναγνωρίζει έξι διαφορετικούς τύπους ορισμού μιας κλάσης:

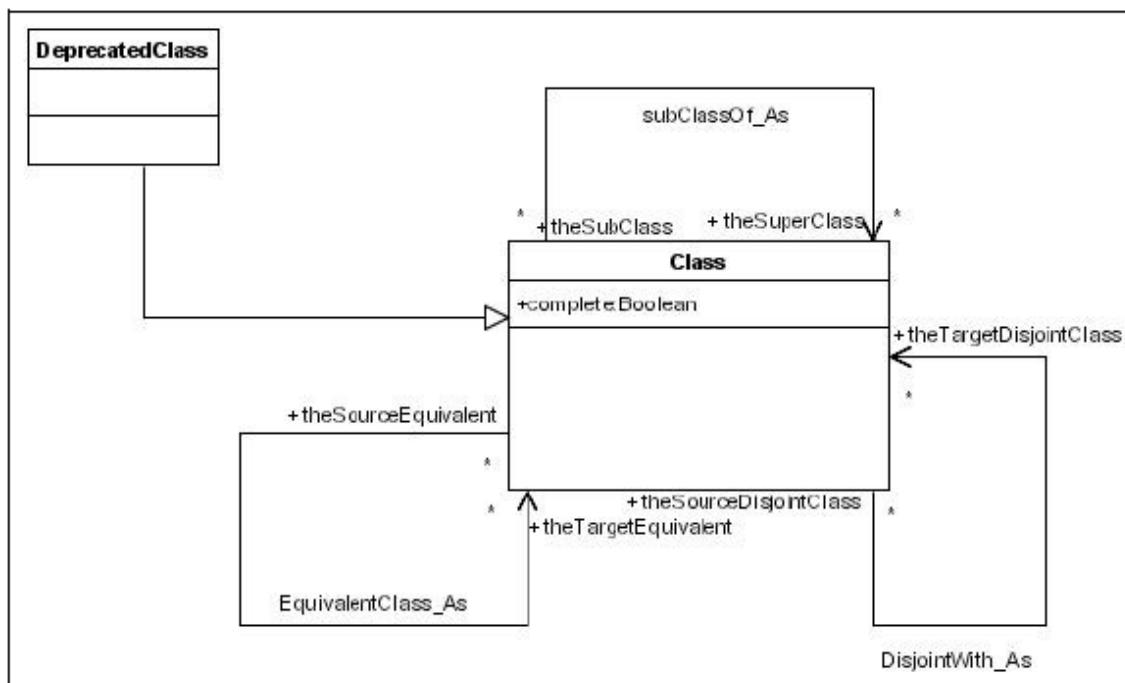
- I. Σαν μια νέα κλάση με κάποιο αναγνωριστικό (δλδ ένα όνομα και κάποιο μοναδικό αναγνωριστικό). Σύμφωνα με αυτόν τον ορισμό μια κλάση ορίζει όλα της τα στιγμιότυπα.
- II. Σαν μια εξαντλητική απαρίθμηση των ατόμων που όλα μαζί αποτελούν τα στιγμιότυπα της κλάσης. Όπως ορίστηκε και παραπάνω μια κλάση μπορεί να οριστεί σαν ένα σύνολο από άτομα. Αυτός ο ορισμός επιτυγχάνεται μέσω της συσχέτισης *oneOf*³⁷ («μία από»). Για παράδειγμα η κλάση “DBEPartner” στην οντολογία του DBE μπορεί να οριστεί ως «μία από» “TUC”, “ISUFI”, “TCD”, “SUN”, “SOLUTA” κ.τ.λ.
- III. Σαν έναν περιορισμό μια ιδιότητας. Αυτήν την περίπτωση θα την αναλύσουμε παρακάτω στην συνέχεια της παραγράφου.
- IV. Σαν την τομή δύο ή περισσότερων ορισμών κλάσεων. Αυτό θα μπορούσε να γίνει αντιληπτό και σαν τον τελεστή της σύζευξης (AND) και πραγματοποιείται με την συσχέτιση της *intersectionOf* (τομή).
- V. Σαν την ένωση δύο ορισμών κλάσεων. Αυτό μπορεί να γίνει αντιληπτό σαν τον ορισμό της διάζευξης (OR). Στο ODM υλοποιείται με τη συσχέτιση του *unionOf* (ένωση).

³⁶ Λέγοντας «άτομα» αναφερόμαστε ουσιαστικά σε στιγμιότυπα της κλάσης δηλαδή σε αντικείμενα των οποίων ο τύπος ταυτίζεται με την συγκεκριμένης κλάσης. Στο εξής οι όροι άτομο και στιγμιότυπο θα πρέπει να θεωρούνται ταυτόσημοι όπου εμφανίζονται στο κείμενο.

³⁷ Σε όλο το κεφάλαιο όπου αναφερόμαστε σε στοιχεία ή αξιώματα του ODM θα χρησιμοποιούμε σκοπίμως πρώτα την αγγλική ονομασία προκειμένου να γίνεται πιο άμεσα η απόδοση της αρχικής σημασιολογίας της λέξης. Επίσης με αυτόν τον τρόπο είναι πιο εύκολο για τον αναγνώστη να κάνει και τη συσχέτιση με τις αντίστοιχες έννοιες της OWL.

επίσημα συσχετίζουν έννοιες). Όπως και η OWL, το ODM ορίζει ειδικές δομές για τον συνδυασμό ορισμών κλάσεων μέσα σε αξιώματα. Συγκεκριμένα ορίζει τις τρεις ακόλουθες δομές:

1. **subClassOf** (υποκλάση): επιτρέπει την έκφραση του γεγονότος ότι η προέκταση μιας κλάσης είναι υποσύνολο της προέκτασης του ορισμού μιας άλλης κλάσης. Η διαφορά σχέση με το αντίστοιχο αξίωμα της OWL είναι ότι στο ODM δεν επιτρέπεται μια κλάση να είναι υποκλάση του εαυτού της.
2. **equivalentClass** (ισοδύναμη κλάση): ουσιαστικά αντικατοπτρίζει ότι ο ορισμός μιας κλάσης έχει την ίδια προέκταση με τον ορισμό μιας άλλης κλάσης. Θα πρέπει να σημειώσουμε ότι κάτι τέτοιο δεν σημαίνει ταυτόχρονα και ισότητα των κλάσεων δηλαδή των εννοιών στις οποίες αναφέρονται.
3. **disjointWith** (αποσύνδεση): υποδηλώνει ότι δύο κλάσεις δεν έχουν καθόλου κοινά μέλη στις προεκτάσεις τους.



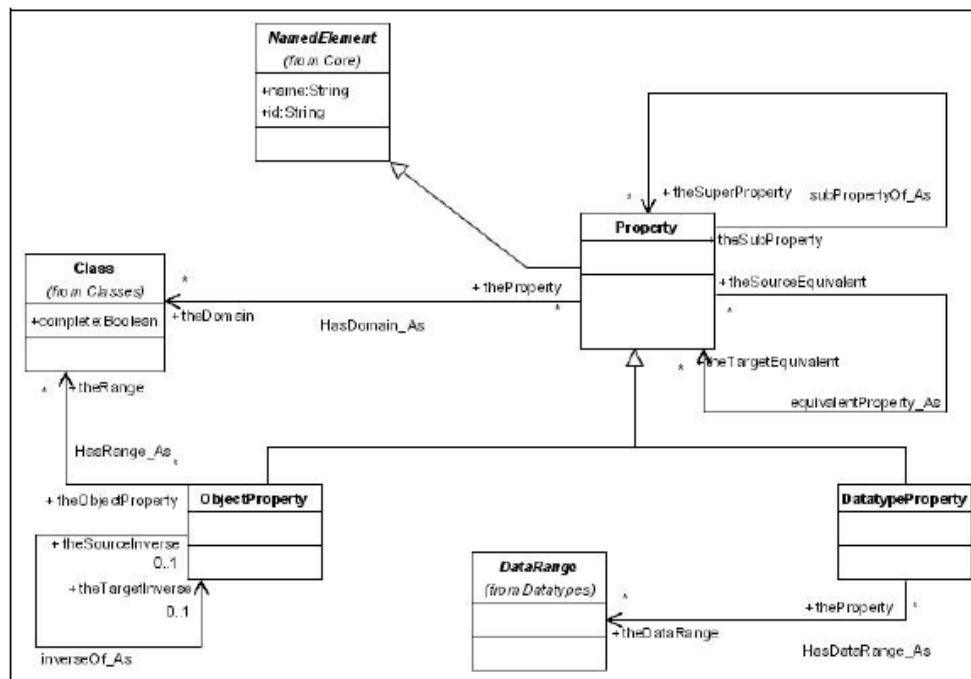
Σχήμα 13: Τα αξιώματα κλάσεων για το ODM

Χαρακτηριστικά Κλάσεων

Μια κλάση χρησιμοποιείται για να περιγράψει μια έννοια όπως η έννοια «άνθρωπος», η έννοια «αυτοκίνητο» κ.ο.κ. σε μια συγκεκριμένη θεματική περιοχή γνώσης. Στον πραγματικό κόσμο η έννοιες έχουν συγκεκριμένα χαρακτηριστικά και

μπορεί να συσχετίζονται με άλλες έννοιες. Για να ικανοποιήσει αυτή την ανάγκη το RDFS χρησιμοποιεί τη βασική έννοια του χαρακτηριστικού. Ένα χαρακτηριστικό μπορεί επίσης να θεωρηθεί και σαν ένας εξελιγμένος μηχανισμός για τη συσχέτιση συγκεκριμένων ατόμων με άλλα άτομα ή με συγκεκριμένες τιμές. Στο παράδειγμα της οντολογίας του DBE το άτομο “BML Editor” που είναι στιγμιότυπο της κλάσης “DBE_Task”, isAssignedTo (έχει ανατεθεί) στο άτομο “TUC” το οποίο με τη σειρά του είναι στιγμιότυπο της κλάσης “DBEPartner”. Σε αυτό το παράδειγμα το κατηγορήμα isAssignedTo έχει οριστεί σαν ένα χαρακτηριστικό του τύπου “DBEPartner” πάνω στην κλάση “DBE_Task”. Ένα στιγμιότυπο αυτού του χαρακτηριστικού συσχετίζει τα δύο άτομα.

Μια κλάση μπορεί να έχει δύο ήδη χαρακτηριστικών: χαρακτηριστικά που συνδέουν στιγμιότυπα μια κλάσης με στιγμιότυπα μιας άλλης κλάσης και χαρακτηριστικά που συνδέουν στιγμιότυπα μιας κλάσης με συγκεκριμένες τιμές δεδομένων. Αν θεωρήσουμε αυτά τα χαρακτηριστικά σαν κατηγορήματα που απλά συνδέουν πεδία³⁸ με αντικείμενα τότε το μόνο που αλλάζει σε αυτά τα δυο είδη χαρακτηριστικών, είναι το αντικείμενο. Παρ’ όλα αυτά το πεδίο δύο χαρακτηριστικών είναι το ίδιο, δηλαδή μια κλάση.



³⁸ Η λέξη πεδίο αναφέρεται σε θεματικά πεδία ή περιοχές από την άποψη του πεδίου ορισμού ενός αντικείμενου και έχει χρησιμοποιηθεί σαν μετάφραση της λέξης domain.

Σχήμα 14: Τα χαρακτηριστικά στο ODM

Σύμφωνα με τα παραπάνω, σχετικά με το αντικείμενο(εύρος³⁹) ενός χαρακτηριστικού ορίζονται οι ακόλουθοι δύο τύποι χαρακτηριστικών:

- I. Χαρακτηριστικά Αντικειμένων. Πρόκειται για χαρακτηριστικά που έχουν σαν αντικείμενο (εύρος) μια άλλη κλάση. Μιλώντας εδώ για το εύρος μιας κλάσης είναι αρκετά ξεκάθαρο ότι αναφερόμαστε στις τιμές που μπορεί αυτή να πάρει με τη διαφορά ότι οι τιμές σε αυτή τη περίπτωση είναι ολόκληρες κλάσεις.
- II. Χαρακτηριστικά τύπων δεδομένων. Πρόκειται για χαρακτηριστικά με συγκεκριμένες τιμές σαν εύρος. Για αυτόν το λόγο χρησιμοποιείται και η έννοια του εύρους δεδομένων (DataRange). Πρόκειται για μια συγκεκριμένη δομή που ορίζει ότι ένα χαρακτηριστικό τύπου δεδομένων έχει είτε συγκεκριμένο τύπο για παράδειγμα «ακέραιος», είτε ένα σύνολο από τιμές δεδομένων.

Τα χαρακτηριστικά στην ODM μπορεί να σχετίζονται μεταξύ τους με σχέσης όπως η ισοδυναμία, η αντιστροφή κ.ά. Επίσης τα χαρακτηριστικά μπορούν να δημιουργήσουν ιεραρχίες μεταξύ τους χρησιμοποιώντας την έννοια του υπο-χαρακτηριστικού. Κάτι ακόμα σημαντικό είναι ότι πάνω στα χαρακτηριστικά ορίζονται περιορισμοί που αφοράνε είτε περιορισμούς πάνω στην πληθυκότητα είτε πάνω στις τιμές των χαρακτηριστικών. Τα αντίστοιχα σχεδιαγράμματα περιλαμβάνονται στο παράρτημα [] και δεν θα μπούμε σε περισσότερες λεπτομέρειες. Σκοπός μας είναι να επικεντρώσουμε στα χαρακτηριστικά εκείνα που ορίζουν τις βασικές σημασιολογικές συσχετίσεις και εκμεταλλευτήκαμε και στη συνέχεια της εργασίας.

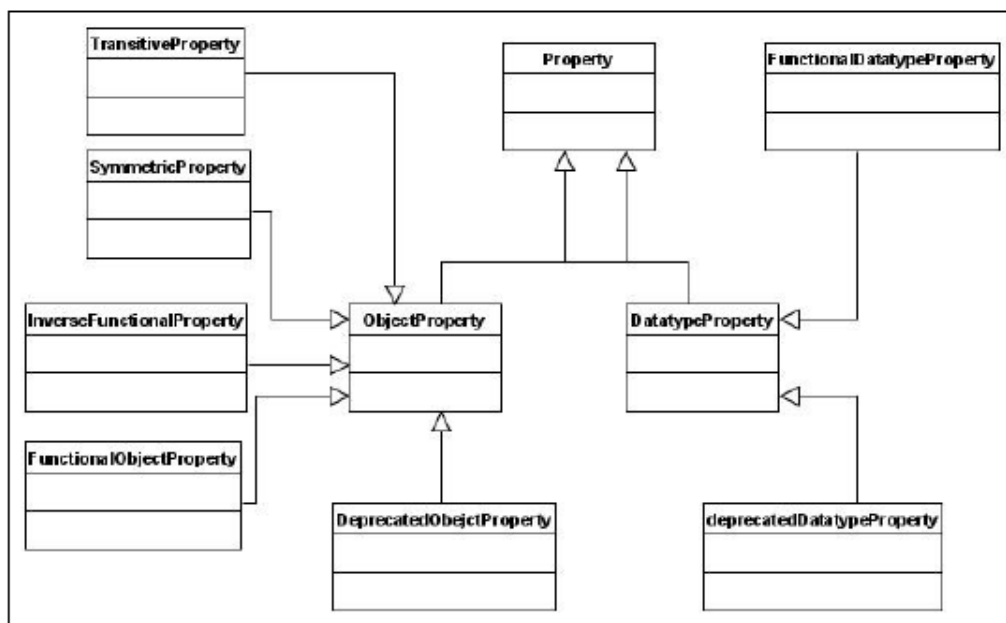
Αξιώματα Χαρακτηριστικών

Ένα αξίωμα χαρακτηριστικού ορίζει τα χαρακτηριστικά ενός χαρακτηριστικού. Στην πιο απλή του μορφή ένα αξίωμα ορίζει την ύπαρξη ενός χαρακτηριστικού, γεγονός που συμβαίνει όταν δημιουργείται ένα στιγμιότυπο της κλάσης Χαρακτηριστικό. Στο ODM ορίζονται οι ακόλουθες δομές για αξιώματα:

39

- Υποχαρακτηριστικό (subPropertyOf)
- Πεδίο και εύρος (όπως τα έχουμε δει, domain και range)
- Ισοδύναμο Χαρακτηριστικό (equivalentProperty)
- Αντίθετο χαρακτηριστικό (inverseProperty)
- Μια σειρά από περιορισμούς πληθυκότητας
- Μια σειρά από λογικά χαρακτηριστικά

Από τα παραπάνω θα θέλαμε να αναφερθούμε στο πεδίο και το εύρος. Μιλώντας για πεδίο αναφερόμαστε στην κλάση για την οποία έχει οριστεί το χαρακτηριστικό ενώ για εύρος, στο εύρος τιμών τις οποίες μπορεί να πάρει το χαρακτηριστικό. Το εύρος ενός χαρακτηριστικού μπορεί να είναι είτε ένας τύπος ή μια απαρίθμηση είτε μια κλάση υποδηλώνοντας ότι ένα στιγμιότυπο της συγκεκριμένης κλάσης μπορεί να είναι τιμή του χαρακτηριστικού. Στο επόμενο σχήμα φαίνονται τα αξιώματα ενός χαρακτηριστικού.



Σχήμα 15: Αξιώματα χαρακτηριστικών στο ODM

Τα επιπλέον στοιχεία που υπάρχουν στο ODM είναι αυτά που επιτρέπουν την προσθήκη σημειώσεων σε μια οντολογία και τέλος η περιγραφή των τύπων δεδομένων. Και τα δύο περιλαμβάνονται στο παράρτημα [].

Στην παρούσα παράγραφο είδαμε συνοπτικά το ODM όπως αυτό αναπτύχθηκε για το DBE στα πλαίσια της παρούσας εργασίας. Ο στόχος όπως είδαμε και στην αρχή ήταν ένα μοντέλο συμβατό με την OWL, την πιο ευρέως αποδεκτή γλώσσα περιγραφής οντολογιών στην κοινότητα του σημασιολογικού ιστού. Με αυτόν τον τρόπο είναι εφικτός και ο χειρισμός υπαρχουσών οντολογιών.

Οι οντολογίες είναι σημαντικό κομμάτι της Βάσης Γνώσης καθώς αποτελούν τον κύριο μηχανισμό για την σημασιολογική διαλειτουργικότητα και την αναπαράσταση της κοινής γνώσης. Οντολογίες για συγκεκριμένες θεματικές περιοχές χρησιμοποιούνται για την περιγραφή κοινής ορολογίας στους διάφορους επιχειρηματικούς τομείς. Οι χρήστες μπορούν επίσης να ορίζουν τις δικές τους οντολογίες. Όλα τα παραπάνω κάνουν φανερή την ανάγκη για ένα ολοκληρωμένο σύστημα προτάσεων για την ανάκτηση πληροφορίας από τη Βάση Γνώσης. Η γλώσσα ερωτήσεων και ο μηχανισμός επεξεργασίας των ερωτήσεων είναι αντικείμενο του επόμενου κεφαλαίου ενώ παράλληλα γίνεται και η πρώτη παρουσίαση του μηχανισμού αυτού στο περιβάλλον του DBE δικτύου εισάγοντας την ανάγκη για προώθηση των ερωτήσεων σε επίπεδο δικτύου.

3.4. Βάση Γνώσης

Η Βάση Γνώσης που αναπτύξαμε όπως είδαμε και στο τέλος της παραγράφου 2.4, έπρεπε να εγκατασταθεί σε κάθε κόμβο του δικτύου και να διαθέτει μια σειρά από χαρακτηριστικά ώστε να υποστηρίζει τις απαιτήσεις του θεωρητικού πλαισίου όπως:

- Να βασίζεται σε πρότυπα: Υποστηρίζει τα πρότυπα του OMG για μοντελοποίηση και ανταλλαγή μοντέλων (πρότυπα για οργάνωση και ανταλλαγή της πληροφορίας, MOF 2.0 – XMI 1.4). Η υιοθέτηση υπαρχόντων προτύπων εξασφάλισε την συμμόρφωση με τις κατευθύνσεις της βιομηχανίας και επέτρεψε τον διαμοιρασμό των μοντέλων μεταξύ διαφορετικών εργαλείων μοντελοποίησης.
- Να είναι επεκτάσιμη. Η Βάση Γνώσης επεκτείνεται εύκολα για να καλύψει άλλα μοντέλα και πηγές πληροφορίας
- Να υποστηρίζει πολλαπλά μεταμοντέλα και μοντέλα. Η Βάση Γνώσης περιλαμβάνει πολλαπλά μεταμοντέλα για οντολογίες, επιχειρήσεις, υπηρεσίες

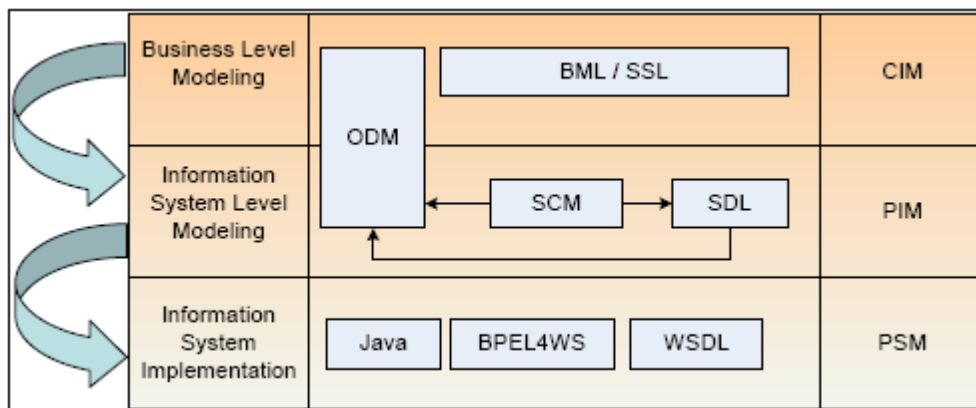
και σύνθεση υπηρεσιών ενώ επιπλέον μεταμοντέλα μπορούν εύκολα να υλοποιηθούν και να υποστηριχθούν.

- Να υποστηρίζει διαφορετικές όψεις των δεδομένων. Υποστηρίζονται διαφορετικές όψεις της πληροφορίας δηλαδή διαφορετικά επίπεδα πρόσβασης στα ίδια μεταδεδομένα με διαφορετικές οπτικές.
- Να είναι ανεξάρτητη πλατφόρμας. Σχεδιασμένη και υλοποιημένη πλήρως για Java, η Βάση Γνώσης είναι ανεξάρτητη πλατφόρμας και έτσι μπορεί να εγκατασταθεί σε οποιαδήποτε επιχείρηση αδιαφορώντας για το οικείο λειτουργικό σύστημα.

Η υποδομή της Βάσης Γνώσης του DBE ακολουθεί την προσέγγιση κατά MOF, για τα μεταδεδομένα. Η Βάση Γνώσης όπως αναπτύχθηκε και στα πλαίσια της παρούσας εργασίας έχει την ικανότητα χειρισμού όλων των τύπων πληροφορίας(δεδομένα και μεταδεδομένα) και εξαγωγής χρήσιμης γνώσης. Η διαχείριση των MOF μεταμοντέλων, μοντέλων και στιγμιότυπων γίνεται μέσω ενός ολοκληρωμένου MOF αποθηκευτικού συστήματος (repository) ενώ για την ανταλλαγή των δεδομένων και των μεταδεδομένων γίνεται χρήση του XML. Από τεχνικής άποψης η αρχιτεκτονική της Βάσης Γνώσης είναι βασισμένη σε ένα συνδυασμό μιας συμβατής με το MOF αποθήκης και ενός συστήματος διαχείρισης δεδομένων. Συγκεκριμένα ενσωματώνεται η αποθήκη μεταδεδομένων της πλατφόρμας NetBeans (NetBeans MDR) (NetBeans MetaData Repository) και σαν επίπεδο διατήρησης των δεδομένων υπάρχει η ευελιξία εκμετάλλευσης διαφορετικών συστημάτων διαχείρισης. Η ευελιξία αυτή στάθηκε χρήσιμη και στην πορεία του προγράμματος όταν αρχικά είχε υποστηριχθεί σχεσιακή βάση Apache Derby ενώ στην συνέχεια χωρίς μεταβολή των πιο πάνω επιπέδων έγινε υιοθέτηση της Berkeley DB XML (The Apache Derby Project) μιας XML δηλαδή βάσης δεδομένων. Αυτό θα φανεί και παρακάτω στο σύστημα επεξεργασίας και παραγωγής των ερωτήσεων. Επίσης η όλη υποδομή υποστηρίζει πλήρως όλη τη σειρά των διαθέσιμων μεταμοντέλων με πρώτο και βασικότερο το Μεταμοντέλο Ορισμού Οντολογιών (**Ontology Definition Metamodel**), αλλά και όλη τη λίστα με τα υπόλοιπα μεταμοντέλα όπως τη Γλώσσα Μοντελοποίησης Επιχειρήσεων (Business Modeling Language), την Γλώσσα Σημασιολογικών Υπηρεσιών (Semantic Service Language), την Γλώσσα Περιγραφής Υπηρεσιών (Semantic Description Language) και το Μεταμοντέλο Σύνθεσης Υπηρεσιών (Semantic Composition Metamodel).

Αυτό που θέλουμε να δείξουμε είναι ότι με την υιοθέτηση του συγκεκριμένου πλαισίου ήταν εφικτή η αντιμετώπιση όλων των αναγκών του συστήματος κάτω από την ίδια τεχνολογική κάλυψη διατηρώντας πάντα την επεκτασιμότητά του. Το πλαίσιο αυτό έπρεπε να χρησιμοποιηθεί και να υποστηριχθεί και από την παρούσα εργασία καθώς κατασκευάστηκαν μια σειρά από υπηρεσίες που απευθύνονταν στην Βάση Γνώσης ενώ το ίδιο το σύστημα των ερωτήσεων έπρεπε να εκμεταλλευτεί τα χαρακτηριστικά που αναλύθηκαν. Τέλος, η Βάση Γνώσης είναι συμβατή με το JMI 1.0 (Java Metadata Interface Specification (v1.0)) το οποίο είναι μια προσπάθεια για την ανάπτυξη ενός πρότυπου Java API για την πρόσβαση και την διαχείριση μεταδεδομένων. Το τελευταίο ίσως είναι απλά μια τεχνική λεπτομέρεια είναι όμως ταυτόχρονα ένα ακόμα πρότυπο που ακολουθήθηκε κατά την υλοποίηση της παρούσας εργασίας τόσο από πλευράς υλοποίησης όσο και από πλευράς μεθοδολογίας.

Στο επόμενο σχήμα υπάρχει μια συνολική άποψη για το πώς γίνεται η εκμετάλλευση της γνώσης του DBE. Η δημιουργία των μοντέλων σε χαμηλότερο επίπεδο του MDA εκμεταλλεύεται τα μοντέλα σε ψηλότερο επίπεδο. Σε επίπεδο επιχείρησης κανείς μπορεί να δει μοντέλα ανεξάρτητα υπολογισμών όπως τα παρουσιάσαμε και στην παράγραφο που περιγράφουν σε πιο πάνω επίπεδο τις μικρομεσαίες επιχειρήσεις. Παραδείγματα τέτοιων μοντέλων είναι η Γλώσσα Μοντελοποίησης Επιχειρήσεων και η Γλώσσα Σημασιολογικών Υπηρεσιών οι οποίες δημιουργούνται με το ODM. Στο επόμενο επίπεδο υπάρχουν στιγμιότυπα του Μεταμοντέλου Σύνθεσης Υπηρεσιών και της Γλώσσας Περιγραφής Υπηρεσιών. Επιπλέον οντολογίες δημιουργούνται με τη βοήθεια του ODM οι οποίες είναι επαναχρησιμοποιήσιμες από τα υπόλοιπα μοντέλα. Με αυτόν τον τρόπο βλέπουν οντολογίες να χρησιμοποιούνται τόσο σε επίπεδο επιχειρήσεων όσο και σε επίπεδο πληροφοριακού συστήματος.



Σχήμα 16: Η MDA προσέγγιση του DBE

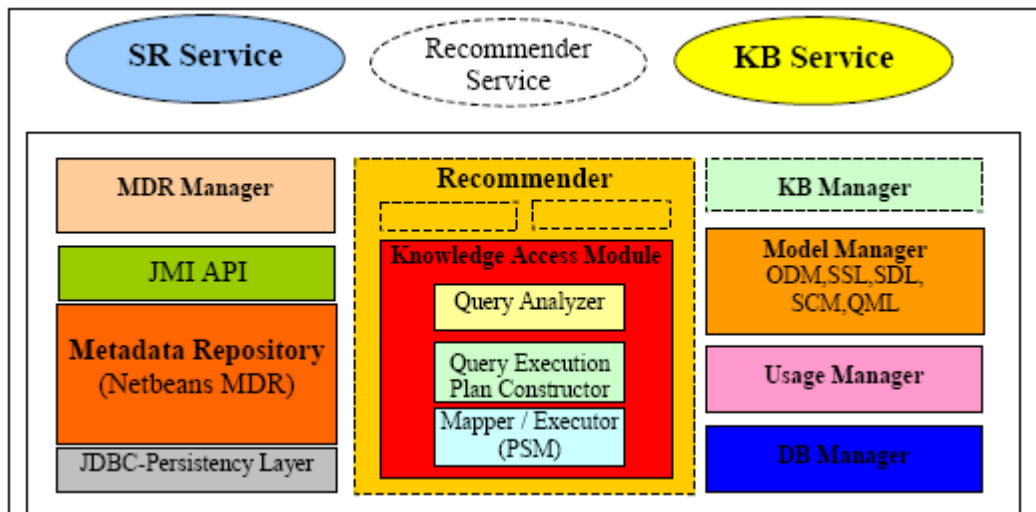
Δεν θα μπορούμε σε περισσότερες λεπτομέρειες στο συγκεκριμένο σημείο γιατί είναι έξω από το αντικείμενο της συγκεκριμένης εργασίας. Παρ όλα αυτά είναι σημαντικό να αναφέρουμε ότι παρόμοια προσέγγιση έγινε για την πρόσβαση στη γνώση. Ορίστηκε μια γλώσσα πρόσβασης – γλώσσα ερωτήσεων – (QML – Query Metamodel Language) η οποία επιτρέπει την ομοιόμορφη και την ανεξάρτητη πλατφόρμας πρόσβαση στην γνώση. Η υλοποίηση της παραπάνω γλώσσας ακολουθεί την MDA προσέγγιση παρέχοντας την δυνατότητα αυτόματης παραγωγής εκτελέσιμου κώδικα δηλαδή ερωτήσεων. Περισσότερες λεπτομέρειες θα δούμε σε επόμενη παράγραφο.

3.5. Μονάδα Πρόσβασης στη Γνώση

Στο Σχήμα 1 Σχήμα 16 παρουσιάζουμε την αρχιτεκτονική του συστήματος που αναπτύχθηκε πάνω από τη Βάση Γνώσης προκειμένου να επιτρέψει την πρόσβαση στη γνώση του συστήματος. Πιο συγκεκριμένα θα δούμε τη μονάδα πρόσβασης στη γνώση (Knowledge Access Module) στην οποία ανήκει τόσο η γλώσσα ερωτήσεων όσο και η επεξεργασία των ερωτήσεων μέχρι την τελική ανάκτηση των δεδομένων ή μεταδεδομένων.

Η Μονάδα Πρόσβασης στη Γνώση αποτελεί βασικό κομμάτι της υποδομής για τη διαχείριση της γνώσης και μέσω μιας διεπαφής ερωτήσεων είναι στη διάθεση των διαφόρων υπηρεσιών ανάλογα το περιβάλλον στο οποίο αναφερόμαστε. Είδαμε εξάλλου από το δεύτερο κεφάλαιο ότι η όλη υποδομή εξωτερικά είναι βασισμένη σε υπηρεσίες. Οι υπηρεσίες στη προκειμένη περίπτωση όπως φαίνονται και στο παρακάτω σχήμα είναι η υπηρεσία της βάσης γνώσης (KB Service) και η υπηρεσία

του σημασιολογικού καταλόγου (SR Service). Περισσότερες λεπτομέρειες για αυτά υπάρχουν στα παραδοτέα του προγράμματος[].



Σχήμα 17: Η μονάδα πρόσβασης στη Γνώση όπως μοντελοποιήθηκε για το DBE σε σχέση με τη Βάση Γνώσης. Φαίνονται τα διάφορα τμήματα της υποδομής που τεχνικά υποστηρίζουν τη διαχείριση της Γνώσης.

Το τμήμα που μας ενδιαφέρει στο παρόν κεφάλαιο είναι αυτό του συστήματος προτάσεων(Recommender) μέσα στο οποίο βρίσκεται η Μονάδα Πρόσβασης στη Γνώση. Αυτή χρησιμοποιεί την QML για την δημιουργία ερωτήσεων και στη συνέχεια ακολουθεί η ανάλυση της ερώτησης, η κατασκευή του πλάνου της ερώτησης και η εκτέλεσή της. Καθώς η QML αποτελεί από μόνη της ένα κομμάτι ανεξάρτητο των πιο κάτω επιπέδων (επεξεργασίας, εκτέλεσης και τελικής βάσης), για να διατηρηθούν τα πλεονεκτήματα των επιλογών σε επίπεδο αρχιτεκτονικής ανάλογη ήταν η αντιμετώπιση και των υπόλοιπων κομματιών. Αυτά αποτελούνται από τα εξής βήματα:

- Ανάλυση της ερώτησης, όπου γίνεται αναγνώριση τόσο των σημασιολογικών στοιχείων των μεταμοντέλων που περιέχονται όσο και των QML εκφράσεων.
- Κατασκευή του πλάνου εκτέλεσης της ερώτησης όπου γίνεται κατασκευή ενός συντακτικού δέντρου της ερώτησης. Το δέντρο αυτό αποτελεί είσοδο για τα υπόλοιπα στάδια της επεξεργασίας και είναι πολύ σημαντικό και ενδιαφέρον το γεγονός ότι παραμένει ανεξάρτητο της μηχανής εκτέλεσης της ερώτησης αλλά και το τελικού συστήματος διαχείρισης της βάσης των δεδομένων.

- Ακολουθεί η παραγωγή του τελικού κώδικα στη κατάλληλη γλώσσα για την εκτέλεση της ερώτησης ανάλογα με το σύστημα αποθήκευσης των δεδομένων. Είσοδος είναι το συντακτικό δέντρο της ερώτησης και γίνεται αναγνώριση των εννοιών του μεταμοντέλου και αντιστοίχισή τους με το σχήμα του συστήματος που αποθηκεύει τα δεδομένα. Πιο συγκεκριμένα η διαδικασία αυτή για την περίπτωση ενός σχεσιακού συστήματος Βάσεων Δεδομένων θα παρήγαγε SQL κώδικα ενώ για την περίπτωση μιας XML Βάσης Δεδομένων θα παρήγαγε XQuery. Τέτοια ανάγκη προέκυψε και κατά τη διάρκεια του προγράμματος αποδεικνύοντας την επεκτασιμότητα, τη διαλειτουργικότητα και την προσαρμοστικότητα του συστήματος.
- Τέλος γίνεται η εκτέλεση της ερώτησης οπότε και ο παραγμένος κώδικας προωθείται στην αντίστοιχη μηχανή εκτέλεσης. Φυσικά, προς το παρόν, αναφερόμαστε σε επεξεργασία σε τοπικό επίπεδο. Εδώ συμπεριλαμβάνεται και η συλλογή των αποτελεσμάτων της ερώτησης.

Από τα παραπάνω βήματα μέσα στα αντικείμενα της εργασίας ήταν το τρίτο με την παραγωγή του τελικού κώδικα και της τελικής ερώτησης και την επεξεργασία της. Η υλοποίηση ενσωματώθηκε στις υπηρεσίες που είναι διαθέσιμες σε κάθε κόμβο μέσω της υποδομής για την πρόσβαση στη γνώση. Με αυτόν τον τρόπο πέρα από την υλοποίηση ήταν πολύ σημαντικό το κομμάτι της επεξεργασίας της ερώτησης αφού ήταν άμεσα συνδεδεμένο με την προώθησή της αλλά και τη συλλογή των αποτελεσμάτων. Από αυτό το σημείο και μετά κύριο στόχο της εργασίας αποτέλεσε η προώθηση της ερώτησης έχοντας βέβαια να αντιμετωπίσουμε την πρόκληση της όσο ποιο αποδοτικής προώθησης της ερώτησης. Στην διαδικασία αυτή, σαν ανάγκη προέκυψε η εκτίμηση του κόστους κάθε ερώτησης όπως είδαμε και στην εισαγωγή του πρώτου κεφαλαίου. Αυτό ήταν ένα ακόμα σημείο συνεισφοράς της παρούσας εργασίας και για αυτό θα τα περιγράψουμε όλα αναλυτικά στο επόμενο κεφάλαιο. Εδώ, θα κλείσουμε με παρουσίαση της επεξεργασίας της ερώτησης και σύντομη αναφορά στην παραγωγή του τελικού κώδικα για την προώθηση και την εκτέλεσή της. Μπορεί κάποια τμήματα να αποτέλεσαν πιο πολύ τεχνικές ανάγκες είναι όμως ουσιαστικά καθώς, αποτέλεσαν ταυτόχρονα απαραίτητο κομμάτι του πλαισίου που έχουμε περιγράψει. Η επεξεργασία μιας ερώτησης και η προώθησή της και έπρεπε και μπόρεσε να εκμεταλλευτεί από τη μια πλευρά όλο το σημασιολογικό πλαίσιο

περιγραφής της γνώσης και από την άλλη πλευρά την υποδομή του συστήματος για την ανάκτηση και μετάδοση της γνώσης.

3.5.1. Επεξεργασία Ερώτησης

Τα κύρια στοιχεία αναφερόμενοι σε ερωτήσεις είναι πρώτον ότι έχουν γραφτεί με τη βοήθεια της QML και δεύτερον ότι αναφέρονται(χρησιμοποιούν) σε τμήματα των οντολογιών που υπάρχουν αποθηκευμένες στο σύστημα. Από τη διαδικασία αυτή θα πρέπει α) να γραφτεί η τελική ερώτηση που θα εκτελεστεί και β) να γίνει ανάλυση της ερώτησης ώστε να είναι εφικτή η εκτίμηση του κόστους της και η αποδοτική προώθησή της.

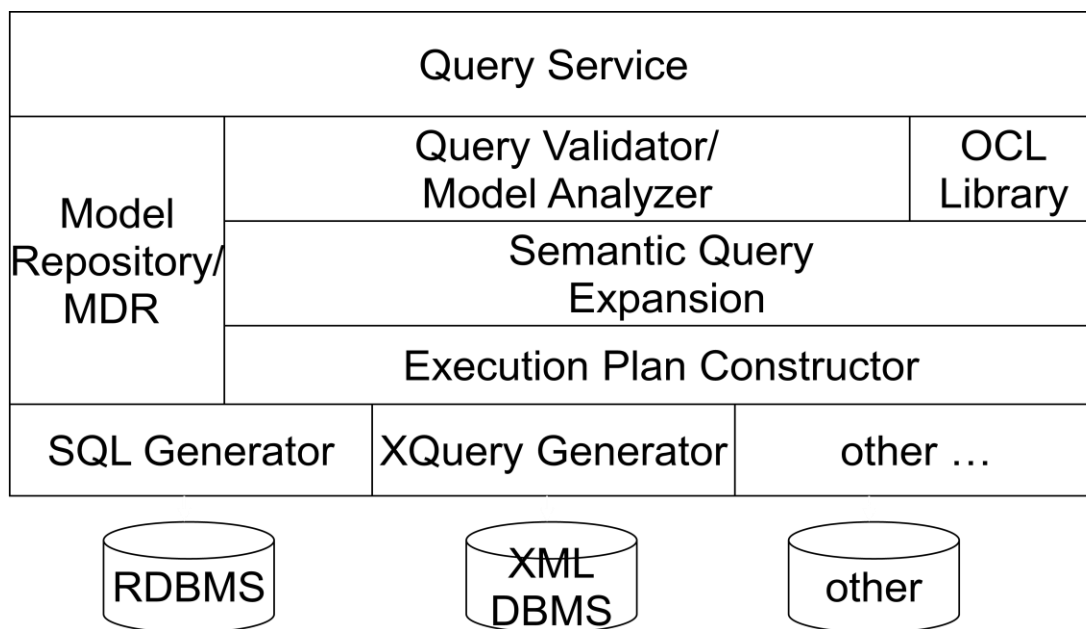
Στην παραγωγή της τελικής ερώτησης εκτός από το να παράγεται ο κώδικας που θα απευθύνεται στο οικείο σύστημα διαχείρισης των δεδομένων θα πρέπει να παράγεται και η τελική μορφή της ερώτησης. Για το δεύτερο δεν θα μπορούμε σε τεχνικές λεπτομέρειες, περιλαμβάνει όμως κάποια σημαντικά στοιχεία όπως:

- η υποστήριξη ασαφών ερωτήσεων σαν μέρος της ανάκτησης πληροφορίας. Μπορούμε να σημειώσουμε ότι την υποστήριξη αυτή την παρέχει και η QML ενώ η διαδικασία ολοκληρωνόταν με την μεταφορά μια τέτοιας απαίτησης σε επίπεδο XQuery παράγοντας τις αντίστοιχες σύνθετες ερωτήσεις. Καλό θα ήταν να αναφέρουμε ότι στο εξής θα θεωρούμε ότι το σύστημα διαχείρισης των δεδομένων ήταν μια XML βάση δεδομένων κάτι που ανταποκρίνεται και στην τελική υλοποίηση.
- η υποστήριξη πολλαπλών οντολογιών. Αυτό πρακτικά δίνει πλήρη ελευθερία κατά τη δημιουργία της ερώτησης σε κάποιον που χρησιμοποιεί το παρεχόμενο πλαίσιο. Τεχνικά, υπάρχει η δυνατότητα αποθήκευσης συνδέσμων και αντιστοιχήσεων μεταξύ των οντολογιών. Από 'κει και περά η επεξεργασία των ερωτήσεων δίνει επεκταμένα δέντρα ερωτήσεων που με τη σειρά τους οδηγούν σε επεκταμένες, ξαναγραμμένες ερωτήσεις που καλύπτουν την ετερογένεια στις περιγραφές της γνώσης. Αυτό φυσικά δεν είναι τόσο χρήσιμο σε τοπικό επίπεδο ή μιλώντας για ένα κόμβο του συστήματος όσο είναι σε επίπεδο δικτύου όπου η ύπαρξη πολλαπλών οντολογιών είναι μια «χρήσιμη» ελευθερία.

Πιο τεχνικά θα δούμε τα παραπάνω και αμέσως, μέσα από την διαδικασία παραγωγής του τελικού κώδικα με απλή μόνο αναφορά σε ότι ήταν έξω από τα αντικείμενα της παρούσας εργασίας.

3.6. Η Μηχανή Εκτέλεσης

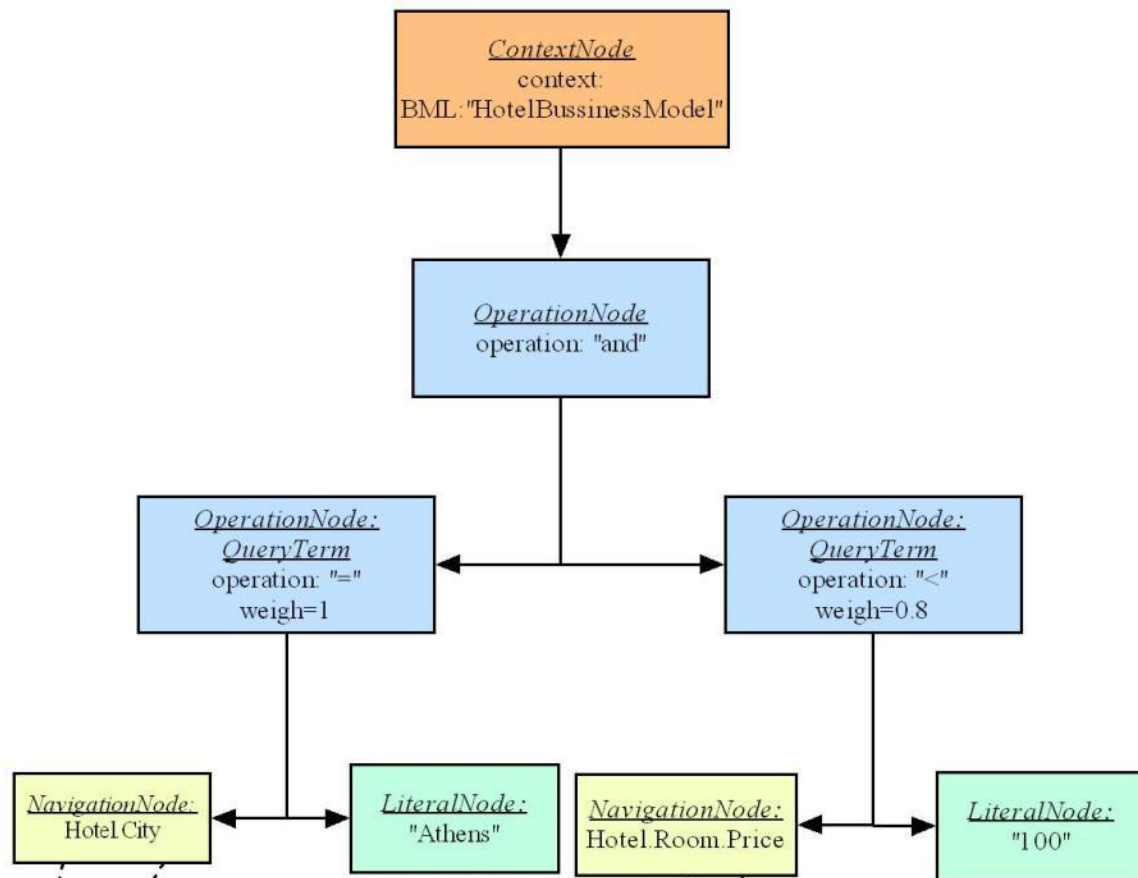
Η QML χρησιμοποιείται για την υποβολή ερωτήσεων σε κάποιο συγκεκριμένο μεταμοντέλο. Ο ρόλος της διαδικασίας παραγωγής της τελικής ερώτησης είναι να δέχεται σαν είσοδο ένα συντακτικό δέντρο και να παράγει τον κώδικα της τελικής ερώτησης και στην κατάλληλη γλώσσα. Αυτό σημαίνει ότι σύμφωνα με την αρχιτεκτονική του συστήματος και προκειμένου να τηρηθεί η ανεξαρτησία του και η επεκτασιμότητά του, η αρχική ερώτηση θα είναι πάντα εκφρασμένη σε QML ενώ τη τελική ερώτηση θα είναι είτε μεταφρασμένη σε SQL είτε σε XQuery είτε σε κάποια άλλη γλώσσα ανάλογα με το σύστημα διαχείρισης των δεδομένων του συστήματος. Αυτό που θα δούμε στη συνέχεια της παραγράφου είναι δυο διαφορετικές περιπτώσεις που χρειάστηκε να καλυφθούν και στα πλαίσια του προγράμματος για να ανταποκριθούν στη σχεσιακή βάση που προέβλεπε αυτό στην αρχή αλλά και στην XML βάση που υιοθετήθηκε τελικά. Στο επόμενο σχήμα βλέπουμε την αρχιτεκτονική της μηχανής εκτέλεσης του συστήματος.



Σχήμα 18: Η αρχιτεκτονική της μηχανής εκτέλεσης των ερωτήσεων του συστήματος

Στα τμήματα που φαίνονται στο Σχήμα 18 θα πρέπει να σημειώσουμε ότι μέχρι την πρώτη κατασκευή του δέντρου εκτέλεσης της ερώτησης, όλα είναι βασισμένα στην QML. Από εκείνο το σημείο και μετά, η επεξεργασία και η εκτέλεση της ερώτησης αποτελούν τον πυρήνα της μηχανής εκτέλεσης που αναλαμβάνει να παράγει την

τελική ερώτηση. Πριν περάσουμε στην περιγραφή της διαδικασίας για κάθε περίπτωση, παραθέτουμε και το συντακτικό δέντρο της QML που είναι αυτό του ακόλουθου σχήματος:



Σχήμα 19: Παράδειγμα Συντακτικού δέντρου για μια QML ερώτηση. Διακρίνουμε τις διαφορετικές περιπτώσεις κόμβων που φέρουν την αντίστοιχη πληροφορία

Με το παραπάνω δέντρο σαν είσοδο ξεκινάνε τα επόμενα στάδια που θα περιγράψουμε στη συνέχεια της παραγράφου.

3.6.1. Παραγωγή τελικής ερώτησης: Η περίπτωση της σχεσιακής Βάσης Δεδομένων

Στα πρώτα στάδια του προγράμματος έγινε προσπάθεια να χρησιμοποιηθεί μια παραδοσιακή, σχεσιακή βάση δεδομένων για όλες τις ανάγκες του συστήματος. Αυτό σήμαινε ότι πρακτικά όλα τα δεδομένα και τα μεταδεδομένα του συστήματος θα έπρεπε να αντιστοιχηθούν σε κάποιο σχεσιακό σχήμα. Αυτή η διαδικασία δεν ήταν αυτόματη και σήμαινε ότι θα έπρεπε να υπάρχει κάποιο συγκεκριμένο σχήμα υλοποιημένο στο διαχειριστή της Βάσης Δεδομένων και από κει και πέρα το

σύστημα να μπορεί να ανακτά κάθε φορά τα ζητούμενα δεδομένα. Αυτό το κομμάτι ήταν και το επόμενο σημείο συνεισφοράς της παρούσας εργασίας.

Συγκεκριμένα σύμφωνα με την αρχιτεκτονική του συστήματος που έχουμε περιγράψει μέχρι τώρα, θα έπρεπε κάποιος χρήστης να μπορεί γνωρίζοντας μόνο κάποιο μεταμοντέλο ή μοντέλο να διατυπώσει οποιαδήποτε ερώτηση σε κάποιον από τους κόμβους και να πάρει πίσω τα αποτελέσματα. Η ερώτηση σε αυτό το στάδιο ήταν μεταφρασμένη με βάση την QML ενώ η τελική βάση ήταν σχεσιακή με αποτέλεσμα η τελική ερώτηση να πρέπει να είναι μεταφρασμένη σε SQL. Η μεθοδολογία που ακολουθήθηκε περιλάμβανε δύο τμήματα:

- α) την αντιστοίχιση των δομών της σχεσιακής βάσης με τις δομές που χρησιμοποιούνται από την QML για τη δημιουργία της ερώτησης
- β) το μηχανισμό δημιουργίας της τελικής SQL ερώτησης

Το πρώτο κομμάτι(α) υλοποιήθηκε με τη μορφή μιας βοηθητικής κλάσης που να περιέχει τις απαραίτητες αντιστοιχίσεις στις οποίες ο μηχανισμός παραγωγής της ερώτησης (β) να μπορεί να ανατρέξει οποιαδήποτε στιγμή και να τις χρησιμοποιήσει. Εδώ θα πρέπει να σημειωθεί ότι η επιλογή της σχεσιακής βάσης για την αποθήκευση και διαχείριση των διαφόρων μοντέλων του προγράμματος αποδείχθηκε όχι ιδιαίτερα ευέλικτη. Για παράδειγμα στο κομμάτι της αντιστοίχισης, ένας μεγάλος αριθμός από σχέσεις και πίνακες έπρεπε να τηρείται με τη μορφή έτοιμου κώδικα(hard coded) προκειμένου να καλυφθούν όλες οι απαραίτητες αντιστοιχίσεις που θα επέτρεπαν τη δημιουργία των ερωτήσεων. Επιπλέον η αρχική δημιουργία του σχεσιακού σχήματος δεν είχε κάποιον προφανή ή αυτόματο τρόπο παραγωγής και δεν ήταν εύκολο να καλύψει σε όλα τα σημεία τις πληροφορίες των αντίστοιχων σχημάτων των διαφόρων μοντέλων. Σε αυτό θα πρέπει να προσθέσουμε την ανάγκη από τον αρχικό σχεδιασμό, το σχήμα να καλύπτει τόσο τα διαφορετικά επίπεδα μοντελοποίησης όσο και την αποθήκευση των τελικών δεδομένων, κάτι που περιέπλεκε το σχεδιασμό του. Επίσης ένα ακόμα πρόβλημα ήταν ότι οποιαδήποτε αλλαγή στο σχήμα επέφερε αλλαγές σε όλα τα επίπεδα ενώ επηρέαζε και τον ίδιο το μηχανισμό παραγωγής των ερωτήσεων αναιρώντας την ανεξαρτησία και την ευελιξία του. Για όλους τους παραπάνω λόγους πολύ γρήγορα αναγνωρίστηκε η ανάγκη για διαφορετική διαχείριση των ημιδομημένων δεδομένων και υιοθετήθηκε μια XML βάση δεδομένων για την οποία θα μιλήσουμε στην αμέσως επόμενη παράγραφο. Αυτό μπορεί να επέφερε κάποιες

Στο επόμενο σχήμα φαίνεται ενδεικτικά το περίπλοκο σχήμα που κατασκευάστηκε για την αποθήκευση του μοντέλου περιγραφής των υπηρεσιών του συστήματος μαζί με ένα μέρος του ODM:



94

γνωστή στο μηχανισμό παραγωγής των ερωτήσεων. Η αντιστοίχιση αυτή περιελάμβανε τόσο πίνακες όσο και σχέσεις και ήταν με τη μορφή βοηθητικών κλάσεων όπως ενδεικτικά:

```
//T12 ClassInstance: ODM
```

```
m_Table t12 = new m_Table("ClassInstance","OntologyClassInstance",
KBInfo.KB_ODM_MODEL_NAME);

t12.addAttribute(new m_Attribute("InstanceID","OntologyClassInstanceID", true));

t12.addAttribute(new m_Attribute("name","name", false));
```

```
//R11 ServiceAttribute - AttrInstance
```

```
m_Relation r11 = new m_Relation("AttrInstance", "AttrType", t3, t10,
"ServiceAttributeID","ServiceAttributeID");

t3.addRelation("AttrInstance", r11);

t10.addRelation("profInstance", r11);
```

Με τον τρόπο αυτό δημιουργήθηκε ένα ευρετήριο σχέσεων και πινάκων το οποίο μπορούσε η παραγωγή των ερωτήσεων να χρησιμοποιήσει σε οποιοδήποτε βήμα. Ανάλογα για παράδειγμα με το μοντέλο που αναγνωρίζεται στην εκάστοτε ερώτηση αναζητείται ο αντιστοιχούμενος πίνακας ώστε να προστεθεί στο σωστό σημείο της παραγόμενης SQL ερώτησης. Καθένας πίνακας διαθέτει επίσης στο ευρετήριό του τα χαρακτηριστικά του και τις σχέσεις στις οποίες συμμετέχει ενώ κρατιούνται και κάποιες πληροφορίες όπως το ποια είναι τα κύρια κλειδιά των σχέσεων. Τέλος όλες οι σχέσεις αποθηκεύονται στις συσχετίσεις μαζί με πληροφορίες για τα άκρα τους με την παρακάτω μορφή:

```
public class m_Relation {

    private AssociationRole roleA;

    private AssociationRole roleB;

    public class AssociationRole{

        private String role;

        private m_Table theTable;
```

private String jAttribute;

.....

Αυτές οι πληροφορίες είναι αρκετές ώστε ο μηχανισμός παραγωγής των ερωτήσεων να γνωρίζει ποιοι πίνακες από το σχεσιακό σχήμα χρειάζονται και επιπλέον ανάλογα και με την ερώτηση του χρήστη ποια θα είναι η τελική ερώτηση ώστε πρώτων να συμπεριληφθούν οι απαραίτητες συνθήκες και δεύτερον να γίνουν οι απαραίτητες συνδέσεις μεταξύ των πινάκων.

Ο Αλγόριθμος

Αλγοριθμικά θα πρέπει να θυμίζουμε ότι είσοδος στον μηχανισμό είναι ένα δέντρο της μορφής του σχήματος 19. Η διαδικασία στη συνέχεια ακολουθεί μια «σε βάθος⁴⁰» διάσχιση του δέντρου με στόχο να αναγνωρίσει τα δομικά του στοιχεία και τη σειρά εμφάνισής τους. Για καθένα από αυτά τα στοιχεία κρατάει τις απαραίτητες πληροφορίες με τη βοήθεια του μηχανισμού αντιστοίχισης ώστε να τις εξάγει στο τελικό στάδιο παραγωγής της SQL ερώτησης και στο σωστό σημείο. Αναλυτικότερα τα στοιχεία που αναγνωρίζονται και οι ενέργειες που αντιστοιχούν σε καθένα από αυτά είναι τα εξής:

1. Κόμβοι λειτουργιών (Operation Nodes)

Πρόκειται για κόμβους που διακρίνονται σε τρεις κατηγορίες

- α) απλοί κόμβοι επιλογής (select) που αναφέρονται στην προβολή των ζητούμενων στοιχείων
- β) λογικοί κόμβοι (boolean) που αναφέρονται σε μια από τις διαθέσιμες πράξεις ένωσης των αποτελεσμάτων (διαζευκτικά, συζευκτικά, με ασάφεια)
- γ) κόμβοι εφαρμογής κάποιας συνθήκης (=,<,> κτλ)

Οι κόμβοι της πρώτης κατηγορίας βρίσκονται πάντα πάνω από κόμβους των άλλων κατηγοριών και προκαλούνε είτε την παραγωγή των τελικών ερωτήσεων είτε την παραγωγή κάποιων σκελών υποερωτήσεων μαζί με τη σχέση που τις συνδέει

2. Κόμβοι πλοήγησης (Navigation Nodes)

⁴⁰ Depth first

Πρόκειται για τους κόμβους που περιέχουν μονοπάτια των ερωτήσεων. Τα μονοπάτια σε επίπεδο μοντέλων είναι κυριολεκτικά μονοπάτια πλοήγησης μέσα στην XML ιεραρχία των μοντέλων. Με τους κόμβους πλοήγησης γίνεται ανάλυση των μονοπατιών στους πίνακες του σχεσιακού σχήματος, κρατιούνται οι αντιστοιχιζόμενοι πίνακες και σχέσεις και τελικά παράγονται οι συνδέσεις(joins) μεταξύ τους που καλύπτουν το αρχικό μονοπάτι

3. Κόμβοι τιμών (Literal Nodes)

Οι τελευταίοι, είναι απλοί κόμβοι που περιέχουν το τελικό κομμάτι κάποιας συνθήκης ανάλογα με τον τύπου που απαιτεί η συνθήκη(συμβολοσειρά, αριθμός κτλ)

Τέλος θα πρέπει να σημειώσουμε ότι χρησιμοποιούνται βοηθητικοί πίνακες για τη συγκέντρωση των αποτελεσμάτων και τη δυνατότητα εφαρμογής πάνω στην παραγωγή τους βαρών και ταξινόμησης εφόσον πρόκειται για κάποια ασαφή ερώτηση.

3.6.2. Παραγωγή τελικής ερώτησης: Η περίπτωση της XML Βάσης Δεδομένων

Όπως είδαμε και νωρίτερα ο ρόλος της διαδικασίας παραγωγής της τελικής ερώτησης είναι να δέχεται σαν είσοδο ένα συντακτικό δέντρο και να παράγει τον κώδικα της τελικής ερώτησης και στην κατάλληλη γλώσσα. Στην τελική υλοποίηση το σύστημα αποθήκευσης ήταν μια XML βάση δεδομένων. Όλα τα μοντέλα, για την μεταφορά, την αποθήκευση και την ερώτησή τους βρίσκονται, όπως είχαμε δει νωρίτερα, σε XMI μορφή και κρατιούνται επίσης στη βάση. Για την απάντηση μιας οποιασδήποτε ερώτηση αυτή θα πρέπει πρώτα να μεταφραστεί σε XQuery. Επιπλέον η χρήση πολλαπλών οντολογιών οδηγεί σε επέκταση της ερώτησης και τελικά σε μια πιο σύνθετη XQuery έκφραση.

Από τεχνικής άποψης για την αναγνώριση των μοντέλων και των οντολογιών κάποιος θα έπρεπε να αναλάβει την αντιμετώπιση των XMI αρχείων με ενιαίο τρόπο ώστε να είναι εφικτή η χρήση τους σαν XML αρχείων χωρίς περιορισμούς. Αυτό επιτεύχθηκε με ένα ανεξάρτητο τμήμα που αναλάμβανε α) να αναγνωρίζει τις αναφορές στοιχείων σε οποιοδήποτε σημείο του εγγράφου⁴¹ και β) την

⁴¹ Πρόκειται για δυνατότητες που δεν παρέχονται από την από την XML και άρα δεν μπορούσαν απευθείας να υποστηριχθούν από την XQuery.

αναγνώριση αναφορών σε στοιχεία από εξωτερικές οντολογίες ή μοντέλα κάτι που οδηγούσε και σε επανεγγραφή των μονοπατιών⁴² μιας ερώτησης.

SQL	XQuery	QML
Relations/tables	XML/trees	Models/graphs
select	return	out
from	for, let	contexts, let
where	XPath	OCL
order	order	sort on collections
group	for – let	contexts – let
Stored procedures	functions	In

Πίνακας 1: Αντιστοίχιση μεταξύ δομικών στοιχείων και εκφράσεων των SQL, XQuery και QML.

Έχοντας τα παραπάνω σαν δεδομένα θα δούμε συνοπτικά την διαδικασία παραγωγής της τελικής ερώτησης. Αρχικά έχοντας θεωρήσει πως είσοδος είναι το συντακτικό δέντρο της ερώτησης, θα πρέπει να γίνει ανάλυσή του. Κατά τη διάρκεια της ανάλυσης και για κάθε κόμβο του δέντρου, παράγονται τα αντίστοιχα τμήματα της τελικής ερώτησης. Η τελική ερώτηση ακολουθεί την γενική μορφή μιας απλής ερώτησης σε XQuery όπως η ακόλουθη:

FOR *list of contexts* **IN** *list of collections*

WHERE *list of conditions*

RETURN *return pattern*

Η παραπάνω μορφή λειτουργεί ως πρότυπο για την παραγωγή οποιασδήποτε ερώτησης και κατά την παραγωγή της τελικής ερώτησης όλα τα τμήματα του παραπάνω προτύπου συμπληρώνονται δυναμικά. Οι συλλογές και το περιεχόμενο στόχος της ερώτησης θα πρέπει να συμπληρωθούν, ενώ θα πρέπει να

⁴² Λέγοντας μονοπάτι και επανεγγραφή της ερώτησης μιλάμε ουσιαστικά με αντικατάσταση της αναφοράς του στοιχείου με τη σωστή διαδρομή του μέσα στη οντολογία που ανήκει. Στόχος είναι να δημιουργηθεί μια σωστή έκφραση(path expression) για την ερώτηση.

σημειώσουμε ότι και υποστηρίζονται και πολλαπλά περιεχόμενα⁴³. Το σημαντικό βέβαια κομμάτι είναι αυτό της λίστας των συνθηκών, οι οποίες θα πρέπει επίσης να εξαχθούν και να προστεθούν σαν μια λίστα από εκφράσεις μονοπατιών(διαδρομών) μαζί με τους περιορισμούς ή τους τελεστές. Τελικά το επιστρεφόμενο αποτέλεσμα έχει τη μορφή ενός XML εγγράφου ανάλογα με τις ανάγκες του ερωτώντα αλλά αυτό δεν θα μας απασχολήσει περισσότερο. Σύμφωνα και με τα προηγούμενα, οι βασικές εργασίες που γίνονται είναι:

- η αναγνώριση των περιεχομένων αναφοράς της ερώτησης
- η κατασκευή όλου του μονοπατιού για κάθε έκφραση που περιλαμβάνεται σε κάθε λειτουργία
- η αναγνώριση και διαχείριση των ίδιων των λειτουργιών. Οι λειτουργίες μπορεί να είναι είτε απλές είτε λογικές

Σε αυτό το σημείο δεν θα μπορούμε σε πολλές λεπτομέρειες. Αξίζει όμως να αναφέρουμε ότι οι μεν απλές λειτουργίες καλύπτονται από το πρότυπο ερώτησης που δώσαμε παραπάνω. Οι λογικές λειτουργίες από την άλλη πλευρά είχαν την ανάγκη ενός πιο σύνθετου προτύπου ώστε από τη μια πλευρά να διατηρείται η γενικότητα του συστήματος και από την άλλη πλευρά να παρέχεται η δυνατότητα για την δόμηση σύνθετων ερωτήσεων με τη χρήση λογικών τελεστών. Επίσης θα θυμηθούμε σε αυτό το σημείο ότι στην QML είχαν γίνει οι απαραίτητες προεκτάσεις ώστε να υποστηρίζονται λειτουργίες Ανάκτησης Πληροφορίας (IR) με τη χρήση ασαφών εκφράσεων. Αυτές θα έπρεπε με κάποιο τρόπο επίσης να υποστηρίζονται για την δημιουργία των τελικών ερωτήσεων τόσο για την ολοκλήρωση του υπόλοιπου πλαισίου όσο και για την τελική απόδοση του συστήματος αφού αυτός ήταν και ο βέλτιστος τρόπος εκτέλεσης των ερωτήσεων. Το γενικό πρότυπο στο οποίο ενσωματώνονται όλες οι παραπάνω ερωτήσεις είναι της ακόλουθης μορφής:

FOR *list of contexts* **IN** *list of collections*

LET

```
$fei := path expression,
$fc1 := path expression [fn:contains(. , value)],
.... ,
$r   := (local:fuzzyor($fei, $fc1, w) + . . ) DIV ws
```

⁴³ Αυτό αναφέρεται στις δυνατότητες που παρέχει η ίδια η XQuery για πολλαπλά περιεχόμενα αναφοράς(multi-context queries)

WHERE *list of conditions*

RETURN *return pattern*

Με μια γρήγορη παρατήρηση μπορεί κανείς να δει ότι έχουν προστεθεί τμήματα για την υποστήριξη σύνθετων ερωτήσεων υπολογισμού ασαφών εκφράσεων ενώ κάποιοι καθαρά μαθηματικοί υπολογισμοί που απαιτούνται γίνονται σε εξωτερικά δηλωμένη συνάρτηση όπως η `fuzzyor()`.

Ο αλγόριθμος

Ο αλγόριθμός που ακολουθείται για να ικανοποιήσει την παραπάνω διαδικασία έχει την ίδια ακριβώς ροή με αυτόν που περιγράψαμε στην προηγούμενη παράγραφο για την περίπτωση της παραγωγής SQL ερωτήσεων. Συνεχίζουν να υπάρχουν οι τρεις κατηγορίες κόμβων που είδαμε στην παράγραφο 3.6.1 αφού σαν είσοδο του συστήματος έχουμε ακριβώς το ίδιο QML δέντρο με την αρχική ερώτηση. Η υπόλοιπη διαδικασία κάνει χρήση των συστατικών της XQuery όπως φαίνονται αυτά στον πίνακα ένα. Η ύπαρξη σύνθετων μονοπατιών μέσα στην ερώτηση μαζί με αναφορές σε διάφορα σημεία ενός εγγράφου και σε άλλες οντολογίες κάνουν τη διαδικασία πιο περίπλοκη. Αναλυτικά για τις τρεις περιπτώσεις κόμβων που αναγνωρίσαμε και πριν, έχουμε την ακόλουθη διαδικασία:

1. Κόμβοι λειτουργιών (Operation Nodes).
2. Κόμβοι πλοήγησης (Navigation Nodes). Είναι η πιο σύνθετη περίπτωση κόμβων και περιέχουν μια λίστα από τμήματα του μονοπατιού πλοήγησης της ερώτησης. Τα τμήματα αυτά βρίσκουν την αντιστοιχία τους σε XPATH εκφράσεις και μπορεί να είναι:
 - i. Κάποιο χαρακτηριστικό (attribute) το οποίο είναι είτε μέρος του μονοπατιού είτε τελικό κομμάτι και προορίζεται για κάποια συνθήκη επιλογής.
 - ii. Ένα κόμβος περιεχομένου αναφοράς (context node) που συνήθως βρίσκεται κάπου στην αρχή και καθορίζει σε ποιο περιεχόμενο αναφέρεται η έκφραση που ακολουθεί
 - iii. Κόμβοι πλοήγησης που είναι και οι πιο σύνθετοι κόμβοι αφού μπορεί να περιλαμβάνουν μια ολόκληρη σειρά από κόμβους σαν τμήμα ενός

μονοπατιού. Οι αντίστοιχες εκφράσεις θα πρέπει να εξαχθούν και να προστεθούν στο υπόλοιπο μονοπάτι.

- iv. Κόμβοι επιλογής. Είναι μια ειδική περίπτωση κόμβων πλοήγησης και μπορεί να περιλαμβάνουν ολόκληρες εκφράσεις είτε με συνθήκες επιλογής είτε ακόμα και με αλγεβρικές εκφράσεις οι οποίες μπορούν και αυτές να είναι τμήμα ενός μονοπατιού.

- 3. Κόμβοι τιμών (Literal Nodes). Είναι η πιο απλή περίπτωση κόμβων και οδηγούν στην προσθήκη ενός αλφαριθμητικού λεξήματος στο τέλος μιας συνθήκης(ισότητας, ανισότητας κ.τ.λ.).

3.7. Υποσύστημα Συστάσεων και Επέκταση Ερωτήσεων

Θα κλείσουμε την παρούσα παράγραφο θυμίζοντας ότι παράλληλα με τα παραπάνω και για την υποστήριξη πολλαπλών οντολογιών κατά τη δημιουργία της ερώτησης, έπρεπε να δίνεται η δυνατότητα αναγνώρισης επεκταμένων συντακτικών δέντρων που στόχος τους είναι να συμπεριλάβουν όλους τους αντίστοιχους όρους των γνωστών στο σύστημα οντολογιών, αυξάνοντας πρακτικά τις πιθανότητες απάντησης μιας ερώτησης. Αυτό με τη σειρά του οδήγησε στην ανάγκη δημιουργίας και επεκταμένων τελικών ερωτήσεων. Φυσικά όλα αυτά θα μπορούσαν να υποστηριχθούν και με τη χρήση απλών ερωτήσεων, τόσων όσοι και οι αντίστοιχοι όροι. Η συλλογή και σύνθεση όμως των αποτελεσμάτων θα ήταν σε αυτή τη περίπτωση πολύ πιο σύνθετη και περίπλοκη. Για αυτόν το λόγο, τελικά δημιουργήθηκε ένα συνολικό πρότυπο ερώτησης που εκτός από τα παραπάνω να περιλαμβάνει και τις επεκταμένες ερωτήσεις.

3.7.1. Υποσύστημα συστάσεων και προφίλ χρηστών

Η υπηρεσία συστάσεων είναι υπεύθυνη για τη διαχείριση των προφίλ των χρηστών. Στα πλαίσια του προγράμματος είχε οριστεί συγκεκριμένο μοντέλο περιγραφής του προφίλ των χρηστών με στόχο να περιγράφει συγκεκριμένες προτιμήσεις τους. Τα έγγραφα αυτά αποθηκεύονται τοπικά για κάθε χρήστη και μπορεί κάθε στιγμή να υπάρχουν περισσότερα του ενός προφίλ για κάποιον χρήστη. Στη συνέχεια η υπηρεσία που περιγράφουμε είναι υπεύθυνη να προωθεί τα προφίλ με τη μορφή ερωτήσεων και να επιστρέφει προτάσεις με τη μορφή

συστάσεων για τον αντίστοιχο χρήστη. Η λειτουργικότητα που είναι διαθέσιμη είναι η ακόλουθη:

- αποθήκευση ενός εγγράφου προφίλ. Κάθε έγγραφο μπορεί να περιέχει περισσότερα του ενός προφίλ για έναν χρήστη και κάθε προφίλ μπορεί να περιέχει πολλαπλές προτιμήσεις
- διαγραφή ενός προφίλ κάποιου χρήστη
- απαρίθμηση των διαθέσιμων προφίλ για έναν χρήστη

Οι παραπάνω λειτουργίες είναι διαθέσιμες σε όλους τους κόμβους του δικτύου και κάθε καινούριο προφίλ είναι στη διάθεση της υπηρεσίας συστάσεων. Η υπηρεσία συστάσεων είναι στη συνέχεια υπεύθυνη να κάνει χρήση των προφίλ και να προσφέρει αυτοματοποιημένα συστάσεις στον χρήστη. Η διαδικασία συλλογής των συστάσεων περιγράφεται αμέσως μετά.

Ο κύριος ρόλος της υπηρεσίας συστάσεων είναι να λειτουργεί σαν αυτόνομος πράκτορας και που να είναι υπεύθυνος για τη συλλογή και παροχή των συστάσεων των χρηστών. Η όλη διαδικασία βασίζεται στην ύπαρξη προτιμήσεων για τους χρήστες με την μορφή προφίλ όπως περιγράφηκε και παραπάνω. Κάθε κόμβος διαθέτει τη δικιά του υπηρεσία συστάσεων εγκατεστημένη. Η υπηρεσία είναι φτιαγμένη ώστε να τρέχει στο υπόβαθρο για όσο χρόνο ο κόμβος είναι ενεργός. Η διαδικασία της συλλογής συστάσεων τρέχει περιοδικά. Φυσικά η περίοδος αυτή μπορεί να μεταβληθεί με βάση κάποια λογική υπόθεση για την ανάγκη συστάσεων αλλά και για την πιθανότητα ύπαρξης νέων. Στη συνέχεια θα περιγράψουμε όλη τη διαδικασία της συλλογής συστάσεων την οποία μπορούμε να χωρίσουμε στα ακόλουθα βήματα:

- i. Ανακάλυψη των τοπικά αποθηκευμένων προφίλ
- ii. Εξαγωγή των περιορισμών που περιγράφονται στα προφίλ
- iii. Δημιουργία QML ερωτήσεων με βάση τα προφίλ
- iv. Προώθηση των ερωτήσεων στο δίκτυο
- v. Συλλογή των αποτελεσμάτων

Η προώθηση των ερωτήσεων στο δίκτυο, βασίζεται στους γενικότερους μηχανισμούς αναζήτησης που αναπτύχθηκαν και περιγράφονται σε επόμενη παράγραφο. Συγκεκριμένα η υπηρεσία συστάσεων κάνει χρήση της υπηρεσίας Σηματολογικού Μητρώου για την ανακάλυψη εμφανιζόμενων υπηρεσιών που να ταιριάζουν με τις προτιμήσεις των χρηστών.

Στην αρχή της διαδικασίας η υπηρεσία αναζητά όλα τα αποθηκευμένα προφίλ. Η λίστα με τα προφίλ είναι απαραίτητη προκειμένου να εντοπιστούν τυχόν αλλαγές από τον τελευταίο συστάσεων. Κάθε χρήστης έχει τη δυνατότητα ανά πάσα στιγμή να προσθέσει ένα καινούριο προφίλ ή να αλλάξει κάποιο υπάρχον. Επίσης κάθε προφίλ συσχετίζεται με ένα συγκεκριμένο μοναδικό αναγνωριστικό του χρήστη στον οποίο ανήκει. Η υπηρεσία δημιουργεί ομάδες ερωτήσεων που αναφέρονται στις προτιμήσεις για κάθε τέτοιο αναγνωριστικό και συλλέγει τα αποτελέσματα για κάθε ομάδα ερωτήσεων.

Το δεύτερο βήμα είναι η εξαγωγή των προτιμήσεων, περιορισμών, που περιέχονται σε κάθε προφίλ και η δημιουργία QML ερωτήσεων που να τις περιέχουν. Κάθε προφίλ μπορεί να περιέχει πολλούς περιορισμούς. Οι περιορισμοί υπάρχουν διατυπωμένοι με τη μορφή OCL εκφράσεων. Σε αυτό το στάδιο της διαδικασίας όλοι οι περιορισμοί συγκεντρώνονται σε μια ομάδα διαζευκτικών ερωτήσεων. Στην πραγματικότητα η προκύπτουσα λίστα ερωτήσεων αντιμετωπίζεται σαν μια λίστα ασαφών ερωτήσεων. Αυτό επιτρέπει στις ερωτήσεις να εφαρμοστούν στη συλλογή των συστάσεων και οι συστάσεις να καταταγούν για το σύνολο του προφίλ. Τέλος για να μπορέσει η υπηρεσία να κάνει χρήση της υπόλοιπης υποδομής του DBE για την προώθηση και συλλογή των αποτελεσμάτων οι τελικές ερωτήσεις είναι γραμμένες σε QML.

Οι QML ερωτήσεις επεξεργάζονται πρώτα τοπικά και στη συνέχεια προωθούνται στο δίκτυο. Ο μηχανισμός προώθησης αναλύεται σε επόμενη παράγραφο και δεν αφορά αυτή καθ' αυτή την υπηρεσία συστάσεων. Στο τελευταίο βήμα της διαδικασίας προκύπτει μια ταξινομημένη λίστα των συστάσεων για κάθε χρήστη. Η λίστα είναι διαθέσιμη μέχρι τον επόμενο κύκλο συστάσεων. Επίσης ο χρήστης έχει τη δυνατότητα ανά πάσα στιγμή να ζητήσει τη λίστα του. Παράλληλα η υπηρεσία συστάσεων λαμβάνει υπόψη δύο σημαντικούς παράγοντες οι οποίοι είναι

Η δυναμική δημιουργία και διαγραφή προφίλ ανά πάσα στιγμή από του χρήστες

Α δυναμική είσοδος νέων κόμβων στο δίκτυο που μπορεί να αντιστοιχεί με τις προτιμήσεις κάποιου χρήστη

Η ανακάλυψη αλλαγών στα αποθηκευμένα προφίλ καλύπτεται στα πρώτα βήματα της διαδικασίας. Η οργάνωση των τελικών συστάσεων καθώς και το γεγονός ότι είναι ταξινομημένες επιτρέπει τυχόν νέα αποτελέσματα να ξεχωρίζουν και να

«διαφημίζονται» στο χρήστη. Η ανακάλυψη ενός νέου κόμβου ή μιας νέας υπηρεσίας εξαρτάται από τους υπόλοιπους μηχανισμούς του DBE.

3.8. Προώθηση Ερωτήσεων

Σε όλο το προηγούμενο κεφάλαιο είδαμε όλο το πλαίσιο για την δημιουργία και επεξεργασία των ερωτήσεων. Περιγράψαμε τη μονάδα ανάκτησης της γνώσης και επικεντρώσαμε στα τμήματα που ήταν αντικείμενο της παρούσας εργασίας. Το σημαντικό στοιχείο είναι ότι σε συνδυασμό με το πλαίσιο που είχε ήδη τεθεί από το δεύτερο κεφάλαιο μπορούμε πλέον να μιλάμε για ένα ολοκληρωμένο σύστημα αποθήκευσης, διαχείρισης και ανάκτησης της γνώσης. Οι στόχοι από την αρχή ήταν η διαλειτουργικότητα και η επεκτασιμότητα του συστήματος και όλα αυτά σε μια προσπάθεια για την δημιουργία ενός ενιαίου πλαισίου που θα μπορούσε να συγκεντρώσει κάτω από την ομπρέλα του διαφορετικούς πελάτες-επιχειρήσεις, από διαφορετικούς τομείς και περιοχές με διαφορετικές ανάγκες περιγραφής των ίδιων, των δεδομένων και των υπηρεσιών τους. Με όσα είδαμε στο παρόν κεφάλαιο, καταφέραμε να έχουμε μια πλήρη εκμετάλλευση της αρχιτεκτονικής και της υποδομής του συστήματος για την επίτευξη των παραπάνω στόχων. Το στοιχείο που λειτουργεί ως κεντρικός άξονας είναι οι οντολογίες ενώ το στοιχείο που λείπει για την ολοκλήρωση του συστήματος είναι η ανάπτυξη των παραπάνω τεχνολογιών σε επίπεδο δικτύου. Η δημιουργία ενός ενιαίου δικτύου ήταν, όπως είχαμε δει, πρωταρχικός στόχος ενώ παράλληλα είχαμε μιλήσει για την επιλογή ενός δικτύου ομότιμων κόμβων. Η υποδομή που χρησιμοποιήσαμε επέτρεπε την δημιουργία του δικτύου σε φυσικό επίπεδο. Η γνώση των φυσικών γειτόνων κάθε κόμβου επιτρέπει απευθείας και την εκμετάλλευση των δικών τους Βάσεων Γνώσης. Σε αυτό το σημείο μιλάμε για την διαδικασία που ονομάζεται γενικά προώθηση ερωτήσεων.

Η προώθηση ερωτήσεων συνήθως αναφέρεται σε κάποιο δίκτυο κόμβων στο οποίο κάθε κόμβος εμφανίζεται ως πιθανώς αποδέκτης της ερώτησης. Όπως και στη δικιά μας περίπτωση η δημιουργία της ερώτησης πραγματοποιείται σε έναν κόμβο του δικτύου ο οποίος θεωρείται και ο αρχικός. Η ερώτηση θα πρέπει να αναλύεται και να εκτελείται σε τοπικό επίπεδο ενώ στη συνέχεια θα πρέπει να προωθείται στους υπόλοιπους κόμβους του δικτύου. Η ανάλυση της ερώτησης σε τοπικό επίπεδο περιλαμβάνει τα βήματα που είδαμε στις προηγούμενες παραγράφους. Η

προώθηση της ερώτησης είναι βασισμένη στην ίδια υποδομή και όπως είχαμε δει και πάνω στην αρχιτεκτονική της μονάδας πρόσβασης στη γνώση, υπάρχουν οι αντίστοιχες υπηρεσίες που αναλαμβάνουν την ίδια την προώθηση. Οι υπηρεσίες αυτές υλοποιήθηκαν στα πλαίσια της παρούσας εργασίας και είναι ικανές να προωθούν ερωτήσεις σε μια λίστα από κόμβους και να συλλέγουν τα αποτελέσματά τους. Σε αυτό το σημείο ανακύπτουν μια σειρά από ερωτήματα, γνωστά και στην βιβλιογραφία, όπως το σε ποιους επιλέγω να προωθήσω μια ερώτηση ή το ποιος είναι ο καλύτερος τρόπος για να προωθήσω μια ερώτηση. Το ποιος κόμβους επιλέγω για την προώθηση της ερώτησης σχετίζεται άμεσα με το υπερκείμενο δίκτυο που είδαμε νωρίτερα ενώ το ποιος είναι ο καλύτερος τρόπος για να προωθήσω μια ερώτηση σχετίζεται εν' γενεί με αυτό που ονομάζουμε κόστος της ερώτησης και εκτίμηση του κόστους μιας ερώτησης. Το ενδιαφέρον είναι ότι στο πλαίσιο στο οποίο βρισκόμαστε η σημασιολογία παίζει τον κυρίαρχο ρόλο. Τα μεν κριτήρια για τη δημιουργία ενός υπερκείμενο δικτύου είναι καθαρά σημασιολογικά η δε εκτίμηση του κόστους των ερωτήσεων μπορεί και αυτή να αντιμετωπιστεί σε επίπεδο σημασιολογικής ανάλυσης. Η ανάλυση αυτή έρχεται σαν αποτέλεσμα της επεξεργασίας που είδαμε στο 3^ο κεφάλαιο αποδεικνύοντας με αυτόν τον τρόπο τη χρησιμότητά της.

Κλείνοντας με το θέμα της προώθησης των ερωτήσεων, από πλευράς υλοποίησης το σύστημα ολοκληρώθηκε εκμεταλλευόμενοι τα φυσικές ιδιότητες του δικτύου και άρα θεωρώντας απλή γνώση των φυσικών γειτόνων κάθε κόμβου. Με αυτό ως δεδομένο η πιο άμεση λύση ήταν η χρήση διάχυσης των ερωτήσεων σε όλο το δίκτυο. Είναι γνωστό ότι αυτή αποτελεί μια σίγουρη πρακτική για την συλλογή όλων των διαθέσιμων αποτελεσμάτων που όμως χαρακτηρίζεται από την αρνητική απόδοση καθώς απαιτεί ένα πολύ μεγάλο αριθμό από μηνύματα που πρέπει να κυκλοφορήσουν στο δίκτυο, μηνύματα που πολλές φορές είναι άσκοπα καθώς απευθύνονται σε κόμβους που δεν διαθέτουν πιθανές απαντήσεις της αρχικής ερώτησης. Παρ' όλα αυτά το πλαίσιο ολοκληρώθηκε και δοκιμάστηκε με τον παραπάνω τρόπο και μάλιστα επιτυχώς καθώς ήταν ολόκληρη η υποδομή που έπρεπε να δοκιμαστεί τόσο από πλευράς δικτύου όσο και από πλευράς των κόμβων για την αποθήκευση, διαχείριση και επεξεργασία όλων των δεδομένων του συστήματος. Από κει και πέρα η υποδομή είναι διαθέσιμη και παρέχει απευθείας τη δυνατότητα εφαρμογής πιο έξυπνων τεχνικών για τη προώθηση των ερωτήσεων.

Εξετάσαμε μια σειρά από ανάλογες τεχνικές και κάναμε κάποιες προτάσεις για την περίπτωση του δικού μας συστήματος. Μέσα από αυτή τη διαδικασία προέκυψε και το ενδιαφέρον θέμα της εκτίμησης του κόστους των ερωτήσεων στο οποίο έγινε περεταίρω έρευνα. Αυτά αποτελούν και το αντικείμενο του επόμενου κεφαλαίου.

Κεφάλαιο 4 - Υπολογισμός Επιλεκτικότητας και Χρήση Σημασιολογικών Συσχετίσεων

4.1. Εισαγωγή: Επιλεκτικότητα και Σημασιολογική Σχετικότητα⁴⁴ σε ODM οντολογίες

Στο προηγούμενο κεφάλαιο μας απασχόλησε μεταξύ των άλλων το θέμα της επεξεργασίας και προώθησης των ερωτήσεων (§ 3.8). Ένα από τα κρίσιμα σημεία στο στάδιο αυτό είναι η εκτίμηση του κόστους κάθε ερώτησης. Για την απάντηση των ερωτήσεων υπάρχει πάντα η πρόκληση της όσο το δυνατόν πιο αποδοτικής προώθησης και επεξεργασίας τους. Η μέτρηση της απόδοσης σε αυτές τις περιπτώσεις βασίζεται στη χρησιμοποίηση όσο το δυνατόν λιγότερων πόρων του συστήματος-δικτύου από τη μια πλευρά και από την άλλη στην όσο το δυνατόν ταχύτερη εκτίμηση της απάντησης. Η φύση των εκάστοτε δεδομένων είναι από τους καθοριστικούς παράγοντες. Από τα παραδοσιακά ακόμα συστήματα βάσεων δεδομένων ο κρίσιμος δείκτης για την απεικόνιση της φύσης των δεδομένων ήταν αυτός της επιλεκτικότητας. Με βάση τα παραπάνω κατά τη διαδικασία εκτίμησης του κόστους μιας ερώτησης θα μπορούσαμε να διακρίνουμε δύο στάδια:

- α) την αναγνώριση συσχετίσεων μεταξύ των δεδομένων και
- β) την κατασκευή συνόψεων, ιστογραμμάτων κ.τ.λ. ή εκτιμήσεων του τελικού κόστους των ερωτήσεων⁴⁵

Η μεθοδολογία που έχει ακολουθηθεί στην παρούσα εργασία καλύπτει και τα δύο στάδια της διαδικασίας είτε με τη μορφή υλοποίησης είτε με τη μορφή προτάσεων. Στο υπόλοιπο όμως της παραγράφου θα ασχοληθούμε με δύο πράγματα:

- α) την αναγνώριση συσχετίσεων μεταξύ των δεδομένων. Οι συσχετίσεις αυτές βασίζονται στη «σημασιολογική σχετικότητα» μεταξύ των δεδομένων μέσω των εξισώσεων που προτείνονται στην παράγραφο 4.3.2.
- β) τον υπολογισμό της επιλεκτικότητας για σύνθετες ερωτήσεις

⁴⁴ Αναφερόμαστε στον όρο semantic relatedness

⁴⁵ Τα πρώτα αναφέρονται σε ενδιάμεσα στάδια επεξεργασίας των δεδομένων για να χρησιμοποιηθούν τελικά στον υπολογισμό των τιμών επιλεκτικότητας είτε απλών είτε σύνθετων εννοιών. Ο υπολογισμός του κόστους των ερωτήσεων αναφέρεται σε υπολογισμό της επιλεκτικότητας της ερώτησης, ανάλογα με την εκάστοτε μεθοδολογία.

Τέλος θα θέλαμε να θυμίσουμε ότι το πλαίσιο εργασίας μας είναι βασισμένο στις οντολογίες ενώ η περιγραφή των δεδομένων στο σύστημά μας γίνεται με τη βοήθεια του ODM. Για το λόγο αυτό έγινε εκμετάλλευση της σημασιολογίας που παρέχει το ODM προκειμένου εντοπιστούν οι συσχετίσεις που υπάρχουν κατά τον ορισμό των δεδομένων και να συμπεριληφθούν στον υπολογισμό του βαθμού). Στο υπόλοιπο του κεφαλαίου θα μιλήσουμε για την έννοια της επιλεκτικότητας γενικά (§ 4.2), σαν τελικό στόχο και θα δούμε πως μπορούμε να θεωρήσουμε δεδομένη την ύπαρξη των τιμών επιλεκτικότητας για μεμονωμένες έννοιες μέσα στα δεδομένα (§4.2.2). Στη συνέχεια θα μιλήσουμε για την αρχή της ανεξαρτησίας (§ 4.2.1) και πως ο τελικός μας στόχος είναι η εκμετάλλευσή της, προτείνοντας μια σειρά εξισώσεων υπολογισμού της επιλεκτικότητας μετά από κατηγοριοποίηση των περιπτώσεων συσχετίσεων των δεδομένων μας με βάση τη σημασιολογία του ODM (§4.1.2). Το κεφάλαιο κλείνει (§4.3.4) με την παρουσίαση μιας σειρά προτεινόμενων μεθοδολογιών για τον υπολογισμό του βαθμού συσχέτισης για συνδυασμούς εννοιών σχετικότητας μεταξύ διαφορετικών στοιχείων της οντολογίας αλλά και της επιλεκτικότητας.

4.1.1. Το πλαίσιο εργασίας

Όπως έγινε σαφές κατά την εισαγωγή στόχος μας ήταν να εκμεταλλευτούμε τα χαρακτηριστικά του συστήματος προκειμένου να μπορούμε από τη μια πλευρά να κάνουμε εκτίμηση της επιλεκτικότητας για σύνθετες ερωτήσεις και από την άλλη πλευρά να μπορούμε να υπολογίσουμε σημασιολογικές συσχετίσεις μεταξύ των εννοιών. Επειδή και για τις δύο περιπτώσεις σημαντικό ρόλο παίζει το πλαίσιο στο οποίο δουλεύουμε και συγκεκριμένα η χρήση του ODM για την περιγραφή των δεδομένων θα θυμηθούμε στη συνέχεια της παραγράφου τα σημεία που μας ενδιαφέρουν. Στο κεφάλαιο 3.3 έγινε μια γενικότερη περιγραφή του ODM. Το γεγονός ότι έχουμε ένα πλαίσιο στο οποίο μιλάμε για οντολογίες και αυτές περιγράφονται επίσημα με χρήση του ODM αποτελεί και το εργαλείο για την επεξεργασία των δεδομένων και την εκμετάλλευση της σημασιολογίας για τους σκοπούς που περιγράψαμε. Για το λόγο αυτό υπάρχει η ανάγκη να περιγράψουμε τα εξής δύο πράγματα: α) τη σημασιολογία του ODM και τα πρωτογενή στοιχεία που προσφέρει για την δημιουργία μιας οντολογίας και β) τα μαθηματικά μοντέλα που θα επιτρέψουν την εκτίμηση της σχετικότητας και της επιλεκτικότητας για

συνδυασμούς από έννοιες. Στο υπόλοιπο της παραγράφου θα κρατήσουμε και θα περιγράψουμε τα στοιχεία του ODM που επιτρέπουν τη δημιουργία συσχετίσεων μεταξύ των εννοιών και στη συνέχεια θα αναλύσουμε το μαθηματικό μοντέλο που προτείνουμε για τους παραπάνω υπολογισμούς εξετάζοντας χωριστά τις δύο περιοχές.

4.1.2. ODM και σημασιολογία

Όπως είδαμε και στο 2^ο κεφάλαιο, το ODM είναι μια γλώσσα για την επεξεργασία του περιεχομένου επιτρέποντας την καλύτερη αυτοματοποιημένη ερμηνεία του. Σαν γλώσσα είδαμε ήδη ότι διαθέτει μια σύνταξη και τα στοιχεία που χρειάζονται για τη κατασκευή οντολογιών. Σε γενικές γραμμές το ODM ορίζει κλάσεις αντικειμένων και αξιώματα πάνω σε αυτές προσδιορίζοντας τα σύνολα στιγμιότυπων των κλάσεων. Κάθε κλάση σχετίζεται με ένα σύνολο που περιέχει τα μέλη της, το οποίο καλείται και προέκταση της κλάσης. Δύο κλάσεις μπορούν να έχουν την ίδια προέκταση αλλά παρ' όλα αυτά να είναι διαφορετικές. Αυτό που μας ενδιαφέρει είναι η περιγραφή των κλάσεων και οι διαφορετικοί τρόποι ορισμού αυτής μέσα από το ODM. Πρόκειται ουσιαστικά για τις σχέσεις που αναζητάμε μεταξύ διαφορετικών εννοιών.

Σύμφωνα λοιπόν και με όσα είδαμε στο κεφάλαιο 3.3 οι σχέσεις του ODM που αντιπροσωπεύουν σχέσεις μεταξύ των εννοιών μια οντολογίας και άρα μας είναι απαραίτητες για να εντοπίσουμε συσχετίσεις μεταξύ των εννοιών που εμφανίζονται σε μια ερώτηση είναι οι ακόλουθες:

- **Υποκλάση**, η οποία ορίζεται απευθείας μεταξύ δυο κλάσεων. Πρακτικά αντιπροσωπεύει την εξειδίκευση μεταξύ εννοιών με την υπέρ-κλάση να έχει τον ρόλο της πιο γενικής έννοιας
- **Ισοδυναμία**, η οποία ορίζεται είτε μεταξύ δύο κλάσεων είτε των περιγραφών δύο κλάσεων και υποδηλώνει την ταύτιση των δύο εννοιών ή μιας έννοιας με μια περιγραφή. Επίσης δύο κλάσεις που είναι ισοδύναμες, διαθέτουν και το ίδιο σύνολο στιγμιότυπων.
- **Αποσύνδεση** μεταξύ δύο κλάσεων ή των περιγραφών δύο κλάσεων. Αντιπροσωπεύει τη σημασιολογική διαφοροποίηση δύο κλάσεων. Υποδηλώνει ότι οι δύο κλάσεις δεν έχουν καθόλου κοινή σημασία

υπαγορεύοντας ότι τα σύνολα των στιγμιότυπων τους δεν έχουν σημείο τομής.

- Εκτός από τα παραπάνω αξιώματα υπάρχει και αυτό της **ιδιότητας** ενός αντικειμένου για να συσχετίσει ένα αντικείμενο με κάποιο άλλο ως χαρακτηριστικό του. Για το σκοπό αυτό σημαντικό ρόλο παίζουν το πεδίο και το εύρος μιας ιδιότητας. Το **πεδίο** μιας ιδιότητας περιορίζει το σύνολο των αντικειμένων στο οποίο μπορεί να εφαρμοστεί η ιδιότητα ενώ το **εύρος** μιας ιδιότητας περιορίζει τις τιμές που μπορεί να πάρει η ιδιότητα.

Πάνω στην ιδιότητα θα πρέπει επίσης να συμπεριλάβουμε την **υποκλάση ιδιότητας** καθώς και την **αντίστροφη ιδιότητα** οι οποίες ορίζονται σαν σχέσεις μεταξύ ιδιοτήτων.

4.2. Επιλεκτικότητα

Στην παράγραφο που ακολουθεί θα ασχοληθούμε με την έννοια της επιλεκτικότητας σαν τη βασική μετρική που απεικονίζει τη φύση των δεδομένων. Επίσης, όπως είδαμε και στην εισαγωγή είναι και το παραδοσιακό μέγεθος από τις βάσεις δεδομένων που χρησιμοποιείται για την εκτίμηση του κόστους των ερωτήσεων και τη βελτιστοποίηση της εκτέλεσής τους.

Ο πιο κλασικός ίσως ορισμός για την επιλεκτικότητα(E) είναι αυτός του κλάσματος του πλήθους των διακριτών τιμών που μπορεί να πάρει μια συγκεκριμένη έννοια μέσα σε μια συλλογή από δεδομένα προς το συνολικό πλήθος των εννοιών που υπάρχουν μέσα σε αυτή τη συλλογή:

$$E = \frac{\text{πλήθος ατόμων για μια συγκεκριμένη έννοια}}{\text{πλήθος ατόμων για όλες τις έννοιες}}$$

Ο υπολογισμός της επιλεκτικότητας αναφέρεται πρακτικά στη διαδικασία εξαγωγής στατιστικών στοιχείων από τα δεδομένα. Τα στατιστικά στοιχεία αφορούνε συνήθως μεμονωμένα χαρακτηριστικά ή έννοιες. Αυτό θα ήταν αρκετό αν οι ερωτήσεις κάθε φορά περιείχαν μόνο μια έννοια, κάτι βέβαια που δεν είναι ρεαλιστικό. Σε ένα πραγματικό σύστημα οι ερωτήσεις που κυκλοφορούνε κάθε στιγμή όχι απλώς περιέχουν ένα σύνολο από έννοιες αλλά ανάλογα και με την

γλώσσα ερωτήσεων συνδυάζονται μεταξύ τους για την ικανοποίηση πιο σύνθετων ερωτήσεων.

Στο συγκεκριμένο πλαίσιο στο οποίο εργαζόμαστε, έχουμε να κάνουμε με δεδομένα περιγεγραμμένα σε ODM. Θεωρούμε δηλαδή την ύπαρξη οντολογιών που βρίσκονται σε ODM. Πάνω σε αυτά τα δεδομένα καλούμαστε να προχωρήσουμε στον υπολογισμό της επιλεκτικότητας για όλες τις έννοιες που περιέχονται σε μια συγκεκριμένη συλλογή. Ο υπολογισμός αυτός γίνεται φυσικά εκ των προτέρων⁴⁶ και οι τιμές αποθηκεύονται σε κάποια κατάλληλη δομή για κάθε μελλοντική χρήση. Σύμφωνα και με τα παραπάνω έχουμε να αντιμετωπίσουμε δύο περιπτώσεις: α) τον υπολογισμό της επιλεκτικότητας για τις μεμονωμένες έννοιες και β) τον υπολογισμό της επιλεκτικότητας για συνδυασμούς εννοιών.

Στη συνέχεια θα μιλήσουμε για τον υπολογισμό της επιλεκτικότητας πράγμα που θα μας βοηθήσει να παρουσιάσουμε την υπόθεση της ανεξαρτησίας. Φυσικά η ανάγκη προκύπτει για την περίπτωση του συνδυασμού εννοιών (β) (όχι για τις μεμονωμένες έννοιες). Για το σκοπό αυτό στη συνέχεια της παραγράφου (§4.2.2) θα δούμε την εκμετάλλευση της σημασιολογίας του ODM για τις περιπτώσεις αυτές. Στην επόμενη παράγραφο (§4.3.4) θα προχωρήσουμε στις προτάσεις για τον υπολογισμό της σημασιολογικής σχετικότητας με βάση το ίδιο πλαίσιο.

Στην περίπτωση του υπολογισμού της επιλεκτικότητας για μεμονωμένες έννοιες, μπορούμε να αντιμετωπίσουμε μια οντολογία σαν μια ιεραρχία από έννοιες ή ακόμα πιο απλά σαν μια συλλογή από έννοιες. Η απλούστευση αυτή είναι αρκετή αν ο στόχος είναι ο υπολογισμός των στατιστικών για μεμονωμένες έννοιες. Σε αυτό το σημείο είναι σημαντικό να σημειώσουμε ότι οποιαδήποτε οντολογία μπορεί να αντιμετωπιστεί σαν ένας γράφος. Μέσα σε αυτόν οποιαδήποτε έννοια μπορεί να αναγνωριστεί με μοναδικό τρόπο κάνοντας χρήση του μονοπατιού μέσα στο γράφο που οδηγεί σε αυτήν και έχοντας σαν δεδομένα ότι κάθε οντολογία διαθέτει τουλάχιστον μια έννοια σαν «ρίζα». Ο υπολογισμός αυτός πραγματοποιείται για όλες τις έννοιες και το αποτέλεσμα είναι της μορφής του επόμενου πίνακα:

⁴⁶ Ο υπολογισμός της επιλεκτικότητας για όλες τις μεμονωμένες έννοιες μιας οντολογίας είναι μια «ακριβή» διαδικασία που όμως έχει αναλυθεί εκτενώς σε σχετικές εργασίες και στη διεθνή βιβλιογραφία.

Έννοια	Επιλεκτικότητα
$\alpha/\beta/\gamma$	0.34
α/β	0.49
$\alpha/\delta/\epsilon$	0.03

Πίνακας 2: Ενδεικτική αντιστοίχιση εννοιών με τη συχνότητα εμφάνισής τους. Στην αριστερή στήλη οι έννοιες σημειώνονται με τη μορφή μονοπατιού που «οδηγεί» σε αυτές

Οι τιμές στη δεξιά στήλη αναφέρονται στη συχνότητα εμφάνισης των εννοιών σαν τιμή της επιλεκτικότητας σύμφωνα με τον κλασικό ορισμό που δώσαμε στην προηγούμενη παράγραφο και κυμαίνονται στο $[0,1]$.

Όπως αναφέραμε και προηγουμένως ένας τέτοιος υπολογισμός είναι αντίστοιχος του υπολογισμού που γίνεται στα παραδοσιακά συστήματα βάσεων δεδομένων. Αυτό που απαιτείται είναι κάποιος μηχανισμός, συνήθως υπολογιστικά ακριβός ο οποίος να μπορεί διατρέχοντας τα δεδομένα να εξάγει τα απαραίτητα στατιστικά στοιχεία. Παρ' όλες τις απαιτήσεις πρόκειται για μια διαδικασία που μπορεί να γίνεται εκ των προτέρων και τα αποτελέσματα να είναι στη διάθεση του συστήματος για κάθε μελλοντική ερώτηση. Τι γίνεται όμως στη περίπτωση που για τις ανάγκες ενός πραγματικού συστήματος έχουμε να αντιμετωπίσουμε την ανάλυση σύνθετων ερωτήσεων που περιέχουν έναν συνδυασμό από έννοιες; Η ανάγκη αυτή μάλιστα υπάρχει ανεξάρτητα από τον τύπο ή τη δομή της ερώτησης. Όταν υπάρχει γνώση των δεδομένων και των ερωτήσεων η συχνότητα εμφάνισης κάποιων συγκεκριμένων συνδυασμών από έννοιες θα μπορούσε να οδηγήσει σε μερική επίλυση του προβλήματος με τον συνυπολογισμό των στατιστικών στοιχείων και για τους όποιους συνδυασμούς από έννοιες εντοπίζονται. Όταν όμως οι συνδυασμοί αυτοί αυξάνονται και ακόμα περισσότερο στη γενική περίπτωση που δε μπορούμε ή δεν έχουμε τη γνώση για να περιοριστούμε σε ένα συγκεκριμένο σύνολο από συνδυασμούς εννοιών, ένας τέτοιος υπολογισμός είναι ανέφικτος γιατί πρακτικά θα σήμαινε τον υπολογισμό της επιλεκτικότητας για όλους τους πιθανούς συνδυασμούς εννοιών που εμφανίζονται σε μια συγκεκριμένη συλλογή δεδομένων.

4.2.1. Η αρχή της ανεξαρτησίας

Η επιλεκτικότητα σαν έννοια αλλά σύμφωνα και με τον παραπάνω ορισμό μπορεί εύκολα να βρει το αντίστοιχό της στη θεωρία πιθανοτήτων. Πιο συγκεκριμένα μιλώντας για επιλεκτικότητα ουσιαστικά αναφερόμαστε στη πιθανότητα εμφάνισης μιας συγκεκριμένης έννοιας μέσα σε μια συλλογή από έννοιες. Αυτός ο ορισμός της επιλεκτικότητας έδωσε στους ειδικούς και την πιο απλή λύση στον τρόπο υπολογισμό της επιλεκτικότητας για συνδυασμό από έννοιες. Με βάση την θεωρία πιθανοτήτων υπάρχει η γνωστή ως «αρχή της ανεξαρτησίας» δύο ή περισσότερων ενδεχομένων. Σύμφωνα με αυτήν όταν για δύο ενδεχόμενα A, B ισχύει ότι:

$$P(B|A) = P(B) \quad \text{Εξ. 1}$$

τότε το ενδεχόμενο B θεωρείται ανεξάρτητο του ενδεχομένου A (στοχαστικά ή στατιστικά ή υπό την έννοια της πιθανότητας). Όπου φυσικά $P(B|A)$ είναι η πιθανότητα εμφάνισης του ενδεχομένου B δοθείσης της εμφάνισης του ενδεχομένου A. Εύκολα αποδεικνύεται⁴⁷ ότι αν ισχύει η παραπάνω σχέση τότε και το ενδεχόμενο A είναι ανεξάρτητο του ενδεχομένου B. Τα προηγούμενα οδηγούνε στη διατύπωση της σχέσης

$$P(A \cap B) = P(A)P(B)$$

ή γενικεύοντας τη σχέση έχουμε ότι για n ανεξάρτητα ενδεχόμενα θα ισχύει η σχέση

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \dots \cdot P(A_n).$$

Αν επιστρέψουμε τώρα στην επιλεκτικότητα και στον υπολογισμό της επιλεκτικότητας για έναν συνδυασμό από έννοιες έχουμε ουσιαστικά να υπολογίσουμε την πιθανότητα εμφάνισης του συνδυασμού των εννοιών με βάση τις γνωστές πιθανότητες για την εμφάνιση των μεμονωμένων εννοιών. Αυτό ίσως είναι και το σημείο που θα θέλαμε να υπογραμμίσουμε: υποθέτοντας ότι οι έννοιες

⁴⁷ Από τη θεωρία πιθανοτήτων έχουμε $P(A|B) = \frac{P(B \cap A)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B)P(A)}{P(B)} = P(A).$

είναι ανεξάρτητες μεταξύ τους μπορούμε εύκολα να προσεγγίσουμε την επιλεκτικότητα για το συνδυασμό δύο εννοιών με τον ακόλουθο τρόπο:

$$E_{1,2} = E_1 * E_2.$$

Η προσέγγιση αυτή είναι αρκετά διαδεδομένη τόσο στο χώρο των βάσεων δεδομένων όσο και στις νέες περιοχές έρευνας που έχουν προκύψει με την εισαγωγή της XML και των οντολογιών σαν τρόπους περιγραφής και αποθήκευσης δεδομένων. Πρόκειται φυσικά και για την υπολογιστικά πιο απλή προσέγγιση για οποιονδήποτε αναλυτή κόστους μιας ερώτησης και για αυτό είναι και αρκετά σημαντική η αξία της.

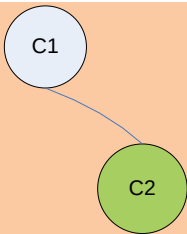
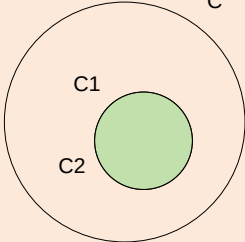
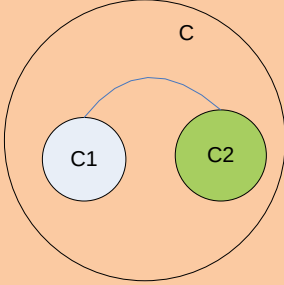
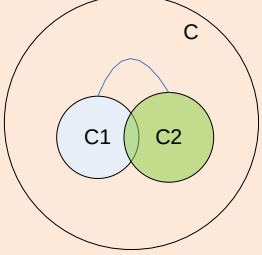
Η έντονη στροφή στη σημασιολογία τα τελευταία χρόνια έβγαλε στην επιφάνεια την διερεύνηση όλου του εννοιολογικού συστήματος πίσω από την περιγραφή των δεδομένων και δημιούργησε την ανάγκη για αποτύπωση τόσο της σημασίας κάθε αντικειμένου όσο και των σχέσεων μεταξύ των αντικειμένων. Η πιο άμεση ίσως συνέπεια ήταν η καθιέρωση της επιστημονικής περιοχής των οντολογιών και της μηχανικής των οντολογιών. Περισσότερα για αυτά αναφέρουμε στα εισαγωγικά κεφάλαια της παρούσας διατριβής. Το σημαντικό στοιχείο είναι μπορέσουμε να επεκτείνουμε την παραπάνω αρχή για τις περιπτώσεις δεδομένων που ανήκουν σε κάποια οντολογία και για τις διαφορετικές περιπτώσεις σημαντικών στοιχείων που υπάρχουν σε αυτές.

4.2.2. Μεθοδολογία

Με βάση όσα είδαμε μέχρι στιγμής θα εξετάσουμε τις διαφορετικές περιπτώσεις εκτίμησης της επιλεκτικότητας κάνοντας χρήση της υπόθεσης της ανεξαρτησίας. Είναι σημαντικό να λάβουμε υπόψη τις υποθέσεις που ισχύουν με βάση το ODM (και την OWL) και οι οποίες συνοψίζονται ως εξής:

1. Οι υποκλάσεις είναι υποσύνολα των αντικειμένων που αντιστοιχούν στην υπέρ-κλάση τύπο-κλάση. Τα υποσύνολα δεν έχουν σχέση γνήσιου υποσυνόλου και επομένως η τομή των συνόλων όλων των υποκλάσεων μπορεί να είναι μικρότερη (ή ίση) του συνόλου της υπέρ-κλάσης
2. Οι κλάσεις μπορεί να είναι ασύνδετες μεταξύ τους
3. Οι κλάσεις μπορεί να είναι ισοδύναμες μεταξύ τους
4. Αν δυο κλάσεις δεν είναι ούτε ασύνδετες ούτε ισοδύναμες τότε θεωρούνται ανεξάρτητες

5. Οι ιδιότητες μπορούν επίσης να αποτελούν υποσύνολο μιας υπέρ-ιδιότητας
Συγκεντρωτικά τα παραπάνω φαίνονται στον επόμενο πίνακα:

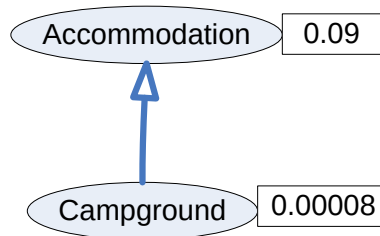
Τοποθεσία	A/A	Περιγραφή	Τύπος	Αναπαράσταση
Α. Στο ίδιο δέντρο, κοινή ρίζα C	A1	Ίδιο path Πατέρας-παιδί	ISA	
	A2	Αδέρφια ισοδύναμα	ISA + equivalent	
	A3	Αδέρφια ασύνδετα	ISA + disjoint	
	A4	Αδέρφια επικαλυπτόμενα	siblings	
	A5	A4 + disjoint		
	A6	A4 + equivalent		
Β. Ιδιότητες	B1	Ιδιότητα	ObjectProperty	

Πίνακας 3

Στη συνέχεια θα δούμε όλες τις περιπτώσεις με τη σειρά εμφάνισης στον πίνακα 3.

A1 Ιεραρχική συσχέτιση (IS-A σχέσεις). Η απλή περίπτωση δύο έννοιες A,B να ανήκουν στην ίδια ιεραρχία και ταυτόχρονα πάνω στο ίδιο μονοπάτι. Πρόκειται για

το παράδειγμα δυο οποιονδήποτε εννοιών που συνδέονται απευθείας με τη σχέση της εξειδίκευσης πατέρας-παιδί ή εγγόνι όπως



Σε αυτήν την περίπτωση είτε η ερώτηση είναι συζευκτική (Accommodation AND Campground) είτε διαζευκτική (Accommodation OR Campground) η επιλεκτικότητα της επηρεάζεται μόνο από το περιεχόμενο της πιο ειδικής έννοιας και για τον υπολογισμό της θα έχω ότι

$$P(\text{Accommodation}, \text{Campground}) = P(\text{Campground})$$

και το πλήθος των αναμενόμενων αποτελεσμάτων θα είναι

$\text{RES}(\text{Accommodation}, \text{Campground}) = N(\text{Campground})$ όπου RES το πλήθος των αποτελεσμάτων της ερώτησης και $N(C)$ το πλήθος των αντικειμένων για την κλάση C.

A2 Συσχέτιση Ισοδυναμίας μέσα στην ιεραρχία. Η περίπτωση της ισοδυναμίας μεταξύ των εννοιών που είναι κομμάτι της ιεραρχίας. Οι έννοιες θα πρέπει να συνδέονται με την σχέση της ισοδυναμίας ανεξάρτητα από το αν είναι ή όχι μέλη της ίδιας ιεραρχίας. Το παράδειγμα θα μπορούσε να είναι το ακόλουθο:



Άλλωστε σύμφωνα και με τον ορισμό του ODM δύο ισοδύναμες έννοιες θα έχουν και ακριβώς το ίδιο σύνολο από στιγμιότυπα και επομένως για τον υπολογισμό της επιλεκτικότητάς τους θα έχω τόσο για την περίπτωση της ερώτησης FamilyDestination AND Destination όσο και για την ερώτηση FamilyDestination OR Destination

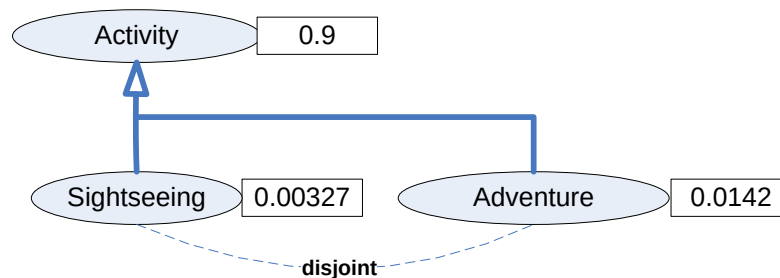
$$P(\text{FamilyDestination}, \text{Destination}) = P(\text{FamilyDestination}) = P(\text{Destination})$$

και το πλήθος των αναμενόμενων αποτελεσμάτων θα είναι

$$\text{RES FamilyDestination}, \text{Destination}) = N(\text{FamilyDestination})$$

= $N(Destination)$ όπου RES το πλήθος των αποτελεσμάτων της ερώτησης και $N(C)$ το πλήθος των αντικειμένων για την κλάση C.

A3 Συσχέτιση ασύνδετων εννοιών μέσα στην ιεραρχία. Η περίπτωση της αποσύνδεσης δύο εννοιών που ανήκουν στην ίδια ιεραρχία. Εδώ έχουμε τη σχέση της αποσύνδεσης(disjoint) όπως στο ακόλουθο παράδειγμα



Εδώ πέρα όπως είδαμε και στην παράγραφο 4.1.2 οι δύο κλάσεις θα πρέπει να μην έχουν κανένα κοινό στοιχείο στα σύνολα με τα στιγμιότυπά τους. Έτσι για τον υπολογισμό της επιλεκτικότητας θα ισχύει:

α) για την ερώτηση Sightseeing AND Adventure

$$P(\text{Sightseeing}, \text{Adventure})=0$$

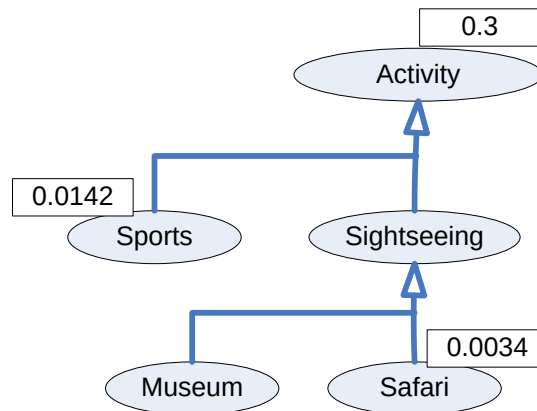
το πλήθος των αναμενόμενων αποτελεσμάτων σε αυτήν την περίπτωση θα είναι $RES(Destination) = 0$

β) για την ερώτηση Sightseeing OR Adventure

$$P(\text{Sightseeing}, \text{Adventure}) = P(\text{Sightseeing}) + P(\text{Adventure})$$

το πλήθος των αναμενόμενων αποτελεσμάτων σε αυτήν την περίπτωση θα είναι $RES(\text{Sightseeing}, \text{Adventure}) = N(\text{Sightseeing}) + N(\text{Adventure})$

A4 Ασυσχέτιστες έννοιες μέσα στην ίδια ιεραρχία. Σε αυτήν την περίπτωση οι έννοιες συσχετίζονται μεταξύ τους μέσω της συγγένειας που έχουν ως μέλη της ίδιας ιεραρχίας και των κοινών στοιχείων που έχουν αποκτήσει από τον κοινό τους πατέρα. Με αυτόν τον τρόπο οι έννοιες έχουν κάποια επικάλυψη μεταξύ τους και μοιράζονται κάποιο κοινό πληροφοριακό περιεχόμενο.



Για τον υπολογισμό της επιλεκτικότητας θα πρέπει να υπολογίσουμε ακριβώς αυτό το κοινό πληροφοριακό περιεχόμενο. Σύμφωνα με όσα είδαμε στην παράγραφο 4.2.1 θα έχουμε ότι

α) για την περίπτωση της ερώτησης Sports AND Safari θα ισχύει

$$P(\text{Sports}, \text{Safari}) = P(\text{Sports}) * P(\text{Safari})$$

β) για την περίπτωση της ερώτησης Sports OR Safari το αναμενόμενο πλήθος των αποτελεσμάτων θα είναι

$$RES(\text{Sports}, \text{Safari}) = N(\text{Sports}) + N(\text{Safari}) - P(\text{Sports}, \text{Safari}) * N(\text{Activity})$$

Όπου $N(C)$ το πλήθος των αντικειμένων που ανήκουν στην κλάση C και RES το πλήθος των αποτελεσμάτων για όλη την ερώτηση.

A5 Υποκλάσεις με ασύνδετες υπέρ-κλάσεις

Πρόκειται για την περίπτωση που οι έννοιες βρίσκονται πιο χαμηλά στην ίδια ιεραρχία αλλά υπάρχουν ασύνδετες υπέρ-κλάσεις σε οποιοδήποτε πιο πάνω επίπεδο. Σε αυτήν την περίπτωση η δύο κλάσεις θα θεωρηθούν ασύνδετες με τη σειρά τους και θα ισχύει ότι και στην περίπτωση A3.

A6 Υποκλάσεις με ισοδύναμες υπέρ-κλάσεις

Πρόκειται για την περίπτωση που οι έννοιες βρίσκονται πιο χαμηλά στην ιεραρχία και υπάρχουν κάποιες ισοδύναμες υπέρ-κλάσεις σε πιο πάνω επίπεδο. Σε αυτήν την περίπτωση θα ισχύει ότι η επιλεκτικότητα για μια ερώτηση $C1$ AND $C2$ θα είναι $N(C1) * N(C2) / N(Clc)^2$, όπου Clc η πιο κοντινή έννοια που αποτελεί άκρο της σχέσης ισοδυναμίας.

B1 Ιδιότητα

Για την επιλεκτικότητα μιας ιδιότητας θα προχωρήσουμε πρώτα στον ορισμό της σαν μέγεθος. Όπως έχουμε δει και κατά την περιγραφή της OWL, οι ιδιότητες ορίζονται με τη βοήθεια ενός χωριστού αντικειμένου αυτού του Property. Επίσης κάθε Property σχετίζεται με ένα πεδίο και ένα εύρος. Ως επιλεκτικότητα ενός $Property_x$ (S_p) ορίζουμε το λόγο

$$S_p(D, R) = \frac{\#Property_x(D, R)}{\#Property_x} \quad \text{Εξ. 2}$$

όπου στον αριθμητή το πλήθος των ιδιοτήτων του συγκεκριμένου τύπου(x) με πεδίο και εύρος τις συγκεκριμένες έννοιες που εμφανίζονται στην ερώτηση και παρονομαστή το πλήθος των ιδιοτήτων του συγκεκριμένου τύπου που εμφανίζονται στην συλλογή ανεξάρτητα από το πεδίο και το εύρος. Στην παρακάτω εικόνα υπάρχει το παράδειγμα της ιδιότητας hasAccommodation. Με βάση το αρχικό παράδειγμα του σχήματος 20 (Σχήμα 23) το hasAccommodation είναι μια ιδιότητα που ορίζεται με πεδίο την έννοια Accommodation και εύρος την έννοια Activity, γενικά. Στο συγκεκριμένο παράδειγμα έχουμε την χρήση της ιδιότητας για δυο συγκεκριμένες υποκλάσεις τις UrbanArea και LuxuryHotel. Επομένως μας ενδιαφέρει να γνωρίζουμε η ιδιότητα hasAccommodation τι επιλεκτικότητα εμφανίζει όχι ανάμεσα σε όλες τις ιδιότητες αλλά συγκεκριμένα για τις ιδιότητες του ίδιου τύπου αφού υπάρχει ελευθερία για το εύρος και το πεδίο που μπορεί να εμφανίζουν.

Παράδειγμα μιας ιδιότητας με δύο έννοιες σαν εύρος και πεδίο είναι το ακόλουθο



Όπου η έννοια UrbanArea είναι στη θέση του πεδίου και η έννοια LuxuryHotel είναι στη θέση του εύρους για την ιδιότητα hasAccommodation και μια υποτιθέμενη ερώτηση θα αναζητούσε “Urban Area AND Luxury Hotel”. Για την επιλεκτικότητα της συγκεκριμένης ιδιότητας μέσα σε όλη τη συλλογή μπορεί η ιδιότητα hasAccommodation να εμφανίζεται 10 φορές. Αντίστοιχα το hasAccommodation για

LuxuryHotel σε UrbanArea μπορεί να εμφανίζεται 3 φορές. Δηλαδή το πληροφοριακό περιεχόμενο της ιδιότητας θα είναι : $P(\text{hasAccommodation}) = \frac{3}{10} = 0.3$. Συνεχίζοντας το θέμα των ιδιοτήτων, όπως είχαμε αναφέρει και αρχικά, το ODM και η OWL δίνουν τη δυνατότητα για δημιουργία ιεραρχίας ιδιοτήτων. Στην περίπτωση αυτή ο υπολογισμός της επιλεκτικότητας για συνδυασμό ιδιοτήτων που αποτελούν μέρος μιας ιεραρχίας θα ακολουθεί την ίδια μεθοδολογία με την περίπτωση A4 όπου κάθε ιδιότητα έχει το ρόλο μια κλάσης, ενός συνόλου αντικειμένων. Στο τέλος του κεφαλαίου θα παραθέσουμε κάποιες δοκιμές που κάναμε για την περίπτωση A4 προκειμένου να δείξουμε την ακρίβεια της μεθοδολογίας στην εκτίμηση της επιλεκτικότητας.

4.3. Σημασιολογική Σχετικότητα

Η επιλεκτικότητα είναι ένα μέτρο για την εκτίμηση του κόστους των ερωτήσεων. Σε πολλές περιπτώσεις όμως οι ερωτήσεις βασίζονται σε κάποιο μοντέλο ή μια οντολογία που δεν είναι κοινή για όλα τα μέλη του δικτύου με αποτέλεσμα να είναι απαραίτητη η συσχέτιση των εννοιών μεταξύ τους. Στην προηγούμενη παράγραφο είδαμε τον υπολογισμό της επιλεκτικότητας των δεδομένων για τις διάφορες περιπτώσεις όταν αυτά είναι περιγεγραμμένα με τη μορφή μιας οντολογίας. Επίσης είδαμε πως κάτι τέτοιο δημιουργεί και φυσικές σημασιολογικές συσχετίσεις που υπάρχουν μεταξύ των δεδομένων. Αυτό που θα μας απασχολήσει στη συνέχεια του κεφαλαίου είναι η εκτίμηση της σημασιολογικής σχετικότητας των δεδομένων κάτω από τις ίδιες συνθήκες.

Στη σχετική βιβλιογραφία υπάρχει αρκετό υλικό με προτεινόμενες μετρικές για τον υπολογισμό σχετικών⁴⁸ αντικειμένων⁴⁹. Με αυτόν σαν στόχο οι περισσότερες μετρικές εφαρμόζονται για τον υπολογισμό σημασιολογικής ομοιότητας⁵⁰ μεταξύ των αντικειμένων. Αυτό που θα πρέπει να γίνει σαφές είναι πως η σημασιολογική ομοιότητα είναι μια ειδική περίπτωση της σημασιολογικής σχετικότητας. Σαν παράδειγμα θα μπορούσαμε να αναφέρουμε ότι οι έννοιες «αυτοκίνητο» και «βενζίνη» μπορεί να μην είναι όμοιες παρ' όλα αυτά μπορούν στα πλαίσια μιας

⁴⁸ Στο εξής όπου χρησιμοποιούμε τον όρο «σχετικός» θα εννοούμε σημασιολογικά σχετικός

⁴⁹ Λέγοντας αντικείμενα αναφερόμαστε σε έννοιες, λέξεις, έγγραφα, θεματικές περιοχές κ.ά.

⁵⁰ δλδ semantic similarity

οντολογίας να παρουσιάζουν υψηλή συσχέτιση (Resnik, Using Information Content to Evaluate Semantic Similarity in a Taxonomy, 1995).

Πιο συγκεκριμένα για να ορίσουμε τη σημασιολογική σχετικότητα θα πρέπει να ξεκινήσουμε από το γεγονός ότι αυτή αφορά δύο όρους των οποίων η σχέση έχει καταγραφεί. Αυτό σημαίνει ότι η σχέση τους παρουσιάζει κάποιο ενδιαφέρον για το συγκεκριμένο πεδίο το οποίο περιγράφουν. Επίσης, μέσα στο συγκεκριμένο πεδίο, η σχέση αυτή, έχει μια μικρή ή μεγαλύτερη πιθανότητα εμφάνισης. Όσο πιο εξειδικευμένη είναι η σχέση (όσο πιο σπάνια εμφανίζεται) τόσο μεγαλύτερο ενδιαφέρον παρουσιάζει. Για παράδειγμα: το ότι ο Γιάννης με τον Κώστα είναι άνθρωποι έχει πολύ λίγο ενδιαφέρον αφού πρόκειται για ένα σχεδόν βέβαιο ενδεχόμενο ($P=1$). Το ότι ο Γιάννης με τον Κώστα είναι αδέλφια παρουσιάζει πολύ μεγαλύτερο ενδιαφέρον και έχει σαφώς λιγότερες πιθανότητες να συμβαίνει. Αυτό ακριβώς προσπαθεί να μετρήσει και η σημασιολογική σχετικότητα σαν μετρική.

Με αυτόν τον τρόπο ενώ όμοιοι όροι σχετίζονται εξαιτίας της ομοιότητάς τους, ανόμοιοι όροι μπορεί να σχετίζονται μέσω άλλων σχέσεων. Για τον παραπάνω λόγο μετρικές για τον υπολογισμό σημασιολογικής ομοιότητας συχνά επεκτείνονται για να καλύψουν τις ανάγκες υπολογισμού σχετικότητας μεταξύ αντικειμένων. Παρ' όλα αυτά ο υπολογισμός ομοιότητας είναι πιο απλός κυρίως γιατί αναφέρεται στην απευθείας σχέση που έχουν δύο αντικείμενα μεταξύ τους. Στην περίπτωση μιας απλής ιεραρχίας ο υπολογισμός του βαθμού ομοιότητας μεταξύ δύο εννοιών μπορεί να προκύψει μόνο με χρήση της σχέσης εξειδίκευσης (IS-A) που υπάρχει μεταξύ τους. Το ίδιο ισχύει για παράδειγμα στην περίπτωση λέξεων όπως στο Wordnet όπου η ομοιότητα μεταξύ δύο λέξεων καλύπτει την έννοια των συνωνύμων ενώ η σχετικότητα θα πρέπει να καλύψει επιπλέον σχέσεις όπως: αντωνυμία, μέρος κ.ά. Προκειμένου να μετρήσει κανείς σημασιολογική σχετικότητα θα πρέπει στους υπολογισμούς να συμπεριλάβει όσες σχέσεις αποτυπώνονται μεταξύ των αντικειμένων. Στην περίπτωση του ODM θα πρέπει να εκμεταλλευτούμε όλες τις σχέσεις που παρέχει το ίδιο. Τις σχέσεις αυτές τις είδαμε και στην παράγραφο 4.1.2. Θα δούμε επίσης τι υπάρχει στην βιβλιογραφία για την περίπτωση της σημασιολογικής ομοιότητας και τέλος θα προτείνουμε την επέκταση των μετρικών για τον υπολογισμό της σημασιολογικής σχετικότητας.

4.3.1. Θεωρία Πληροφορίας και Υπολογισμός Ομοιότητας

Μέχρι αυτό το σημείο έχουμε διαπιστώσει την ανάγκη για τον υπολογισμό της σχετικότητας για συνδυασμούς εννοιών. Επίσης έχουμε παρουσιάσει το πλαίσιο περιγραφής των εννοιών με τη βοήθεια του ODM και στην προηγούμενη παράγραφο είδαμε τις σημασιολογικές δυνατότητες που παρέχει το ODM για συσχετισμό των εννοιών μεταξύ τους. Επιστρέφοντας στον αρχικό στόχο, το πρόβλημα που ανακύπτει είναι αυτό του υπολογισμού της σχετικότητας για συνδυασμό εννοιών με βάση τις τιμές επιλεκτικότητας για τις μεμονωμένες έννοιες. Όπως έχουμε ήδη δει η επιλεκτικότητα σαν έννοια ταυτίζεται με αυτήν της πιθανότητας εμφάνισης της συγκεκριμένης έννοιας μέσα στην συλλογή στην οποία ανήκει.

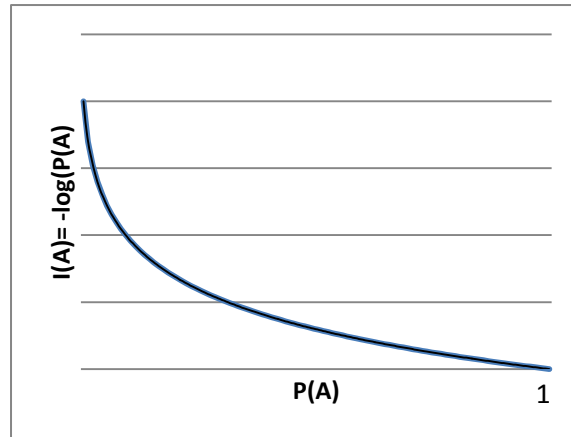
Αυτό που απομένει είναι να βρεθεί ένας επίσημος τρόπος για τον υπολογισμό αυτού του βαθμού συσχέτισης. Σε αυτό το σημείο θα ανατρέξουμε στα εργαλεία που μας παρέχει η Θεωρία Πληροφορίας.

Η Θεωρία Πληροφορίας έχει ασχοληθεί εκτεταμένα με τον υπολογισμό του πληροφοριακού περιεχομένου των εννοιών. Μιλώντας για συσχέτιση δύο εννοιών αναφερόμαστε στη σημασιολογική τους ομοιότητα χωρίς να μας ενδιαφέρει το σύνολο των αντικειμένων που τις αντιπροσωπεύουν. Όταν για παράδειγμα συγκρίνουμε την έννοια «ποτάμι» με την έννοια «αυλάκι» δεν μας ενδιαφέρει να συγκρίνουμε το σύνολο από ποτάμια με το σύνολο από αυλάκια. Μας ενδιαφέρει να συγκρίνουμε την γενική έννοια «ποτάμι» με την γενική έννοια «αυλάκι». Για να μπορέσουμε να εκμεταλλευτούμε τις εργασίες που έχουν προκύψει από την περιοχή της θεωρίας πληροφοριών θα πρέπει πρώτα απ' όλα να ορίσουμε αυτό που προηγουμένως ονομάσαμε πληροφοριακό περιεχόμενο. Το πληροφοριακό περιεχόμενο (Ross, 1997(1976)) μιας έννοιας ορίζεται με βάση την πιθανότητα εμφάνισης της έννοιας στη συλλογή ή στο έγγραφο στο οποίο ανήκει με χρήση της έννοιας της εντροπίας από τον Shannon (Shannon, 1948). Πιο συγκεκριμένα

$$I(A) = -\log P(A)$$

όπου A ένα συγκεκριμένο γεγονός και $P(A)$ η πιθανότητα εμφάνισής του. Δηλαδή όσο μεγαλύτερη η πιθανότητα εμφάνισης μιας έννοιας τόσο μικρότερο το πληροφοριακό περιεχόμενο το οποίο φέρει. Με απλά λόγια, όσο πιο κοινή είναι μια

έννοια τόσο λιγότερη πληροφορία μας παρέχει για το περιεχόμενο που χαρακτηρίζει.



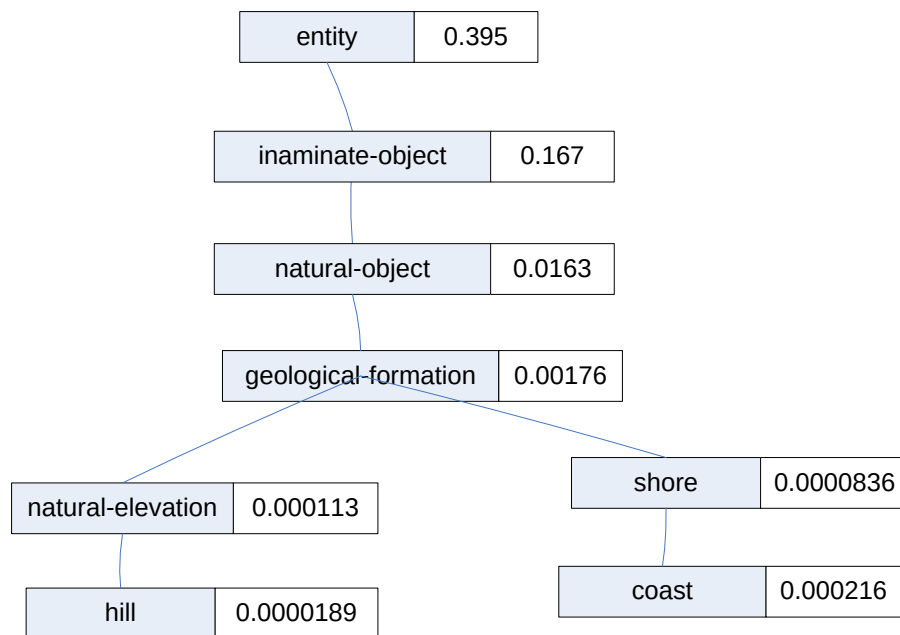
Σχήμα 21: Σχέση πιθανότητας εμφάνισης μιας έννοιας με το πληροφοριακό περιεχόμενο που την χαρακτηρίζει. Στο οριζόντιο άξονα, όσο αυξάνει η πιθανότητα εμφάνισης $P(A)$ μιας έννοιας τόσο μειώνεται το πληροφοριακό περιεχόμενο το οποίο φέρει. Όταν η εμφάνιση της έννοιας είναι βέβαιη ($P(A)=1$) το πληροφοριακό περιεχόμενο που φέρει είναι μηδενικό⁵¹.

Με αυτόν τρόπο μπορούμε εύκολα να υπολογίσουμε το πληροφοριακό περιεχόμενο για οποιαδήποτε έννοια. Ο παραπάνω ορισμός ταιριάζει και διαισθητικά με το τι θα έπρεπε να είναι μέτρο της πληροφορίας (Cover & Thomas, 1991). Το επόμενο βήμα είναι, πώς δοθέντος του πληροφοριακού περιεχομένου δυο εννοιών μπορούμε να βρούμε το κοινό πληροφοριακό τους περιεχόμενο με άλλα λόγια το ζητούμενο βαθμό συσχέτισής τους. Σε αυτό το σημείο θα πρέπει να παραθέσουμε τη σχετική βιβλιογραφία.

Στη συνέχεια της παραγράφου θα δούμε μια πρώτη εφαρμογή της έννοιας του πληροφοριακού περιεχομένου για την περίπτωση ιεραρχίας εννοιών με σχέσεις εξειδίκευσης (ISA). Ήδη από το 1993 ο P. Resnik [(Resnik, Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, 1998), (Resnik, Using Information Content to Evaluate Semantic Similarity in a Taxonomy, 1995)] είχε ασχοληθεί με το θέμα του υπολογισμού της σημασιολογικής ομοιότητας μέσα σε IS-A ιεραρχίες στηριζόμενος στην έννοια του πληροφοριακού περιεχομένου. Για κάθε έννοια c μέσα σε μια ιεραρχία, ονομάζει $P(c)$ την πιθανότητα εμφάνισης στιγμιότυπων της συγκεκριμένης έννοιας μέσα στην ιεραρχία. Σύμφωνα με όσα θέσαμε και παραπάνω

⁵¹ Όταν, για παράδειγμα, μια οντολογία διαθέτει μοναδική ρίζα R τότε η πιθανότητα εμφάνισής της είναι $P(R)=1$. Όταν οι ρίζες είναι περισσότερες R_i θα πρέπει $\sum P(R_i) = 1$.

όσο η πιθανότητα αυξάνεται τόσο μειώνεται το πληροφοριακό περιεχόμενο της έννοιας και επίσης όσο πιο αφηρημένη είναι μια έννοια τόσο μικρότερο πληροφοριακό περιεχόμενο φέρει. Σε αυτό το σημείο μπορούμε να δούμε το ενδεικτικό παράδειγμα τις παρακάτω οντολογίας. Πρόκειται για κομμάτι μέσα από το Wordnet:



Σχήμα 22: Τμήμα της ιεραρχίας του Wordnet. Στο σχήμα εμφανίζονται οι έννοιες μαζί με τη συχνότητα εμφάνισής του.

Σύμφωνα με τον Resnik δύο έννοιες είναι τόσο σχετικές όσο η πληροφορία που μοιράζονται από κοινού. Η πληροφορία που μοιράζονται δύο έννοιες ορίζεται από το πληροφοριακό περιεχόμενο των εννοιών στις οποίες υπάγονται μέσα στην ιεραρχία. Πιο επίσημα όρισε ότι

$$\text{sim}(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log(p(c))]$$

όπου $S(c_1, c_2)$ είναι το σύνολο των εννοιών στις οποίες υπάγονται οι c_1 και c_2 . Τελικά τη μέγιστη(max) πληροφορία θα περιέχει ο ελάχιστος κοινός τους πατέρας. Ενδεικτικά με βάση την ιεραρχία του σχήματος(Σχήμα 22) ισχύει το ακόλουθο παράδειγμα:

$$\begin{aligned} \text{sim}(\text{shore}, \text{hill}) &= \max_{c \in \{\text{entity}, \text{inanimate-object}, \text{natural-object}, \text{geological-information}\}} [-\log(p(c))] = \\ &= -\log(P(\text{geological-information})) = -\log(0.00176) \end{aligned}$$

Ο D. Lin (Lin, 1998) βασιζόμενος στην δουλειά του Resnik, προσπάθησε να δώσει μια σειρά από ορισμούς, κατευθύνσεις και εφαρμογές του παραπάνω σημασιολογικού μέτρου ομοιότητας. Το αποτέλεσμα ήταν η παρακάτω βελτιωμένη μαθηματική φόρμουλα:

$$sim(X_1, X_2) = \frac{2 * \log P(C_0)}{\log P(C_1) + \log P(C_2)} \quad \text{Εξ. 3}$$

θεωρώντας ότι αναφερόμαστε πάλι σε μια ιεραρχία από έννοιες και ότι $x_1 \in C_1$, $x_2 \in C_2$ και C_0 είναι η πιο εξειδικευμένη κοινή έννοια στην οποία υπάγονται οι x_1 , x_2 , με άλλα λόγια ο πιο κοντινός τους κοινός πατέρας. Ακολούθησαν και άλλες εργασίες (Formica, 2007) προς την κατεύθυνση των παραπάνω κάνοντας εκτιμήσεις και αξιολογήσεις παρόμοιων μέτρων για την συσχέτιση εννοιών μεταξύ τους με κοινό όμως στοιχείο το πλαίσιο της ιεραρχίας εννοιών. Και ενώ η χρήση ιεραρχιών από έννοιες είναι πολύ διαδεδομένη, στο πλαίσιο που έχουμε θέσει για την παρούσα εργασία έχουμε να αντιμετωπίσουμε οντολογίες μέσα στη γενικότερη αύξηση της χρήσης οντολογιών μετά την επαναχρησιμοποίηση της έννοιας από τις περιοχές της πληροφορικής. Μια ιεραρχία μπορεί να χαρακτηριστεί και ως κομμάτι μιας οντολογίας ή μια οντολογία από μόνη της, μιας και η πιο ισχυρή και σημαντική σχέση είναι όντως αυτή της γενίκευσης (εξειδίκευσης). Με την εξάπλωση όμως των οντολογιών και την ύπαρξη γλωσσών σαν την OWL ήταν φανερό ότι η όλη περιοχή προέκυψε από την ανάγκη αποτύπωσης σχέσεων μεταξύ των εννοιών της οντολογίας πέρα από την βασική σχέση της γενίκευσης (εξειδίκευσης). Τέτοιες είναι και οι σχέσεις του ODM που αναλύσαμε στη προηγούμενη παράγραφο και είναι σημαντικές όταν προσπαθούμε να αναγνωρίσουμε συσχετίσεις και ομοιότητες μεταξύ εννοιών που ανήκουν στην ίδια οντολογία. Αυτό που θα δούμε στην επόμενη παράγραφο είναι πως με τη βοήθεια της εισαγωγής των εννοιών από την περιοχή της θεωρίας πληροφοριών θα οδηγηθούμε στον υπολογισμό τιμών σχετικότητας για συνδυασμούς εννοιών με βάση τις σχέσεις στις οποίες ανήκουν.

4.3.2. ODM και Υπολογισμός Σχετικότητας

Σε αυτή την παράγραφο θα περιγράψουμε πως με την βοήθεια των προσεγγίσεων που είδαμε προηγουμένως, θα προχωρήσουμε στον υπολογισμό της σημασιολογικής σχετικότητας για τις διαφορετικές περιπτώσεις συσχετίσεων εννοιών που μπορεί να εμφανιστούν κατά την ανάλυση μιας ερώτησης. Θα πρέπει να διευκρινίσουμε ότι προς το παρόν αναφερόμαστε σε συνδυασμούς εννοιών θεωρώντας πρακτικά ζεύγη εννοιών που συμμετέχουν σε ερωτήσεις είτε σαν συζεύξεις είτε σαν διαζεύξεις.

Οι εργασίες που είδαμε στην παράγραφο 4.3.1 εκμεταλλεύονται τη γνώση της τιμής της επιλεκτικότητας για κάθε έννοια και με βάση τι έννοιες που είδαμε από τη θεωρία πληροφoρίας προχωράνε στον υπολογισμό της σχετικότητας των εννοιών για τις περιπτώσεις ιεραρχιών εννοιών. Στην παράγραφο που ακολουθεί θα μιλήσουμε για την επέκταση των παραπάνω εννοιών ώστε να συμπεριλάβουμε τις έννοιες που μας ενδιαφέρουν για την περίπτωση μια οντολογίας με τις σχέσεις που υποστηρίζει το ODM.

4.3.3. Υπολογισμός Σημασιολογικής Ομοιότητας

Πριν προχωρήσουμε είναι σημαντικό να αναλύσουμε τη θεωρητική βάση των προτάσεων που θα ακολουθήσουν με βάση το υπόβαθρο που είδαμε στην προηγούμενη παράγραφο. Η πρώτη και σημαντικότερη πρόταση που μας ενδιαφέρει είναι ότι χρησιμοποιώντας την έννοια του πληροφοριακού περιεχομένου, η ομοιότητα δύο εννοιών δίνεται από τον ακόλουθο τύπο:

$$sim(A, B) = \frac{IC(common(A, B))}{IC(description(A, B))}$$

όπου $IC(common(A, B))$ η ποσότητα της πληροφορίας που μοιράζονται από κοινού οι δύο έννοιες και $IC(description(A, B))$ η ποσότητα πληροφορίας που περιγράφει και τις δύο έννοιες. Ο παραπάνω τύπος οδήγησε και στην εξίσωση 2 του Lin που είδαμε προηγουμένως. Η εξίσωση 2 θα ήταν αρκετή αν αναφερόμασταν μόνο σε περιπτώσεις ιεραρχιών εννοιών. Αυτό που λείπει είναι επέκταση της προσέγγισης ώστε να συμπεριληφθούν και άλλου είδους σχέσεις. Τελικά θα πρέπει να προκύψει μια μεθοδολογία που θα δίνει τη δυνατότητα να υπολογίζουμε τη σχετικότητα δύο εννοιών με κάποια από τις υπόλοιπες σχέσεις που μας παρέχει η OWL.

Σημασιολογική Απόσταση και Αμοιβαιότητα

Σε αυτό το σημείο θα πρέπει να προσφέρουμε και την επέκταση της εξίσωσης 1 για την περίπτωση που δύο έννοιες δεν είναι ανεξάρτητες και τον υπολογισμό εκ νέου με βάση αυτής του πληροφοριακού περιεχομένου δύο εννοιών. Συγκεκριμένα η πιθανότητα εμφάνισης μιας έννοιας, δοθείσης της εμφάνισης μιας δεύτερης έννοιας από την οποία εξαρτάται, θα προκύπτει ως εξής:

$$P(c|p) = \frac{P(c \cap p)}{P(p)} = \frac{P(c)}{P(p)}$$

Θέλοντας τώρα να υπολογίσουμε το πληροφοριακό περιεχόμενο του συνδέσμου μεταξύ των δύο εννοιών, στην πιο απλή δηλαδή περίπτωση, εύκολα καταλήγει κανείς⁵² στο ότι:

$$IC(c, p) = -\log(P(c|p)) = IC(c) - IC(p) \text{ Εξ. 4}$$

Το αποτέλεσμα της εξίσωσης 3 αντιστοιχεί πρακτικά στην διαφορά μεταξύ των τιμών πληροφοριακού περιεχομένου των δύο αλληλοεξαρτώμενων εννοιών. Αυτή η τιμή μας δίνει τη σημασιολογική απόσταση μεταξύ των δύο εννοιών και αντιστοιχεί πρακτικά στο βάρος που έχει η σύνδεση μεταξύ τους. Αν μια έννοια είναι πολύ πιο γενική από κάποια άλλη η απόστασή τους θα είναι και μεγαλύτερη. Σε κάποιες από τις περιπτώσεις που θα δούμε παρακάτω είναι πιο εύκολος ο απευθείας υπολογισμός της σημασιολογικής απόστασης από αυτόν της σημασιολογικής σχετικότητας. Αυτό συμβαίνει φυσικά γιατί όπως αναφέραμε και προηγουμένως δεν είναι όλες οι σχέσεις που θα χρησιμοποιήσουμε ιεραρχικές. Πρακτικά αυτό δεν αλλάζει τον τελικό υπολογισμό της σημασιολογικής σχετικότητας αφού εύκολα μπορεί κανείς να μετατρέψει τη σημασιολογική απόσταση σε σχετικότητα, όπως θα δούμε στην επόμενη παράγραφο.

Στον υπολογισμό της συσχέτισης των εννοιών θα συμπεριλάβουμε ένα ακόμα μέτρο για τις περιπτώσεις εκείνες που οι έννοιες δεν βρίσκονται μέσα σε κοινή ιεραρχία και η ερώτηση του χρήστη αναφέρεται στις ίδιες τις έννοιες αγνοώντας τυχόν άλλη συσχέτιση μεταξύ τους. Σε αυτήν την περίπτωση θα λάβουμε υπόψη την αμοιβαιότητα των εννοιών δηλαδή το κατά πόσο μία ή περισσότερες ιδιότητες συσχετίζουν τις δύο έννοιες και ταυτόχρονα κατά πόσο για κάποιες από αυτές τις ιδιότητες ισχύει και η αντίστροφη έννοια. Πιο συγκεκριμένα όσον αφορά τις ιδιότητες που δύο έννοιες μοιράζονται από κοινού θα έχουμε:

⁵² $\log(A/B) = \log A - \log B$

$$\text{για } \sum_{i=1}^n p_{i1} > 0, \text{ } rel_{op1}(C_1, C_2) = \frac{\sum_{i=1}^n p_{i12}}{\sum_{i=1}^n p_{i1}} \text{ Εξ. 5}$$

όπου p_{i1} υποδηλώνει τη σχέση μεταξύ C_1 και C_i και p_{i12} υποδηλώνει ότι και οι δύο έννοιες C_1, C_2 σχετίζονται με την έννοια C_i . Η βασική ιδέα είναι πως σχετικές έννοιες θα μοιράζονται και περισσότερες κοινές ιδιότητες. Ταυτόχρονα αν εξετάσουμε τον αριθμό από τις κοινές ιδιότητες των εννοιών που διαθέτουν και αντίστροφη έννοια μπορούμε σταθμίσουμε τις έννοιες με τις πιο ισχυρές συσχετίσεις. Για τις αντίστροφες ιδιότητες θα έχουμε:

$$\text{για } \sum_{i=1}^n p_{i1} > 0, \text{ } rel_{op2}(C_1, C_2) = \frac{\sum_{i=1}^n p_{invi12}}{\sum_{i=1}^n p_{i1}} \text{ Εξ. 6}$$

όπου p_{invi12} υποδηλώνει ότι και οι δύο έννοιες συσχετίζονται και αντίστροφα. Ανακεφαλαιώνοντας η βασική ιδέα είναι ότι όσες περισσότερες κοινές ιδιότητες έχουν δύο έννοιες τόσο πιο πολύ συσχετίζονται. Επίσης όσο περισσότερες από τις ιδιότητες έχουν και την αντίστροφή τους τόσο πιο ισχυρές είναι. Τα δύο παραπάνω μέτρα θα συνδυάσουμε αμέσως παρακάτω σε ένα συνολικό και θα χρησιμοποιήσουμε έναν παράγοντα ώστε να μπορούμε να ενισχύσουμε το καθένα από τα δύο ανάλογα με την εφαρμογή. Συνδυάζοντας τα δυο παραπάνω μέτρα θα έχουμε για τη συσχέτιση των εννοιών:

$$sr_{op}(C_1, C_2) = \left(f_1 \frac{\sum_{i=1}^n p_{i12}}{\sum_{i=1}^n p_{i1}} \right) + \left(f_2 \frac{\sum_{i=1}^n p_{invi12}}{\sum_{i=1}^n p_{i1}} \right) \text{ Εξ. 7}$$

Με $f_1, f_2 \geq 0$ και $f_1 + f_2 = 1$.

Στην απλή περίπτωση και αυτό που θα θεωρήσουμε για τις δικές μας ανάγκες θα έχουμε $f_1 = f_2$. Φυσικά στις περιπτώσεις που δεν υπάρχουν αντίστροφες ιδιότητες θα κάνουμε χρήση μόνο της μιας από τις δυο μετρικές.

4.3.4. Μεθοδολογία

Με όσα είδαμε από την αρχή της παραγράφου 4.4 έχουμε καταλήξει σε δύο μετρικές που πρέπει να εκμεταλλευτούμε για να καταλήξουμε στον υπολογισμό της σημασιολογικής σχετικότητας (sr^{53}). Θα πρέπει να διακρίνουμε τις εξής δύο περιπτώσεις:

⁵³ semantic relatedness

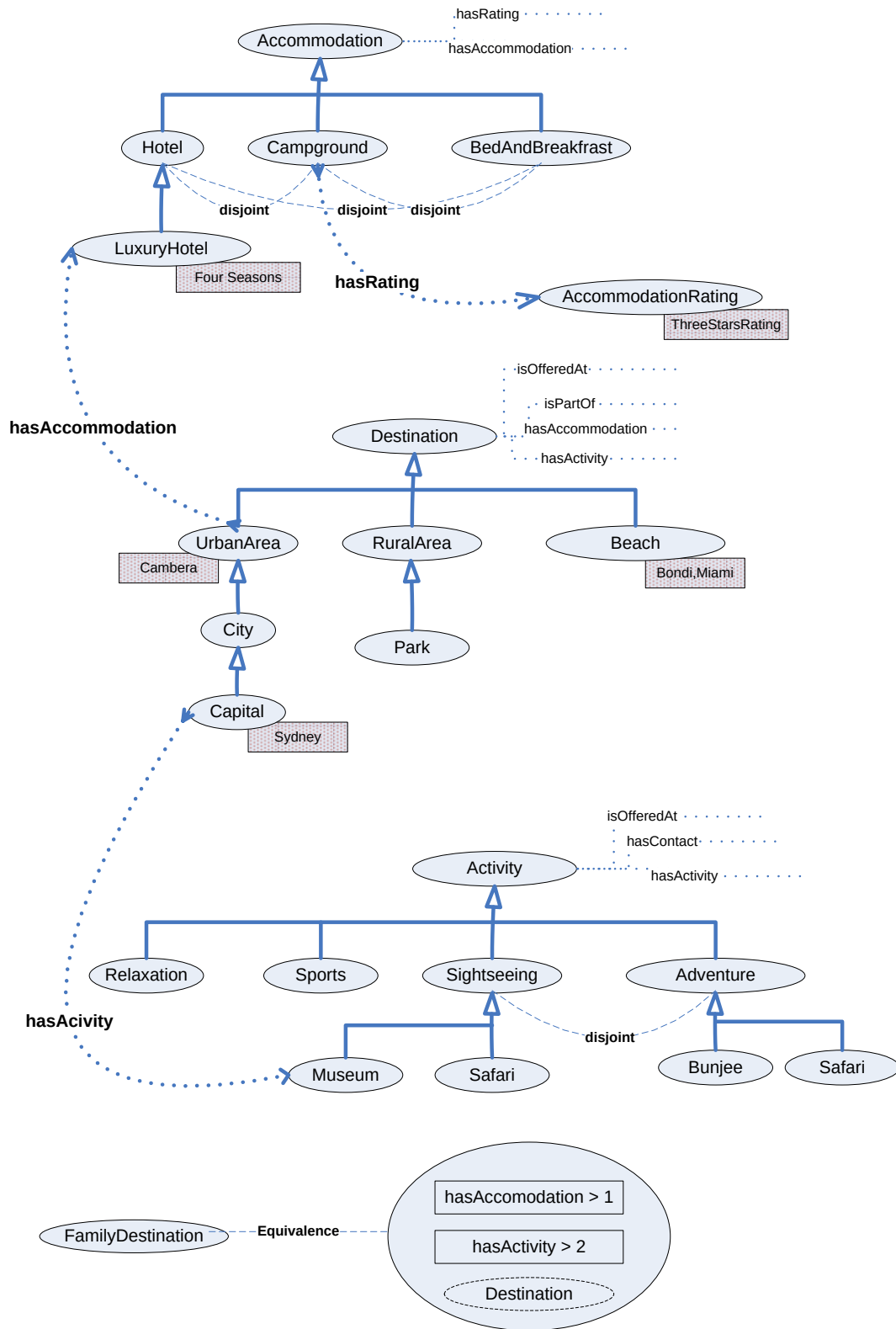
- i. να πρόκειται για ιεραρχία και να είναι εφικτός ο απευθείας υπολογισμός της σημασιολογικής σχετικότητας μέσω της ομοιότητας των κόμβων με χρήση της εξίσωσης 2. Σε αυτήν την περίπτωση $sr(A,B) = sim(A,B)$
- ii. να είναι εφικτός μόνο ο υπολογισμός της σημασιολογικής απόστασης οπότε απαιτείται η μετατροπή του σε σημασιολογική σχετικότητα. Το μέγεθος μπορεί εύκολα να υποστεί μια γραμμική μετατροπή χωρίς να επηρεαστεί το αποτέλεσμα και τελικά να έχουμε $sr(A,B) = dis_{max} - sd(A,B)$, όπου dis_{max} η μέγιστη σημασιολογική απόσταση που εμφανίζεται στη συλλογή μας.

Σε ότι θα δούμε στη συνέχεια θα αναφερόμαστε άλλοτε σε σημασιολογική ομοιότητα και άλλοτε σε σημασιολογική απόσταση ανάλογα με την περίπτωση σχέσης στην οποία αναφερόμαστε.

Για να είναι πιο εύκολη η κατανόηση των περιπτώσεων που θα περιγράψουμε θα χρησιμοποιήσουμε στο εξής το παράδειγμα μιας ενδεικτικής οντολογίας. Η οντολογία φαίνεται στο ακόλουθο σχήμα (Σχήμα 23) και αποτελεί τμήμα μιας πραγματικής οντολογίας του τομέα των ταξιδιών. Στο σχήμα 20 παρουσιάζουμε απλά το κομμάτι της οντολογίας με τους κόμβους που μας ενδιαφέρουν. Στις περιπτώσεις που ακολουθούν απομονώνουμε κάθε φορά το τμήμα που μας ενδιαφέρει και προσθέτουμε τιμές επιλεκτικότητας στους κόμβους ώστε να μπορούμε να δώσουμε και αριθμητικά παραδείγματα. Θα πρέπει να θυμίσουμε ότι η επιλεκτικότητα κάθε κόμβου έχει υπολογιστεί εκ των προτέρων με βάση τη συχνότητα εμφάνισης της αντίστοιχης έννοιας μέσα στη συλλογή και αυτό σαν εφαρμογή του ορισμού για την επιλεκτικότητα που είδαμε στην αρχή του κεφαλαίου. Για να είναι σαφές και για να είμαστε και συνεπείς με την έννοια της πιθανότητας εμφάνισης μιας έννοιας που χρησιμοποιήσαμε σαν βάση των υπολογισμών μας (βλ. πληροφοριακό περιεχόμενο), ισχύει

$$P(C) = \frac{freq(c)}{N}$$

όπου $freq(c)$ η συχνότητα εμφάνισης της έννοιας C εκτιμώμενη με βάση τη συχνότητα των ατόμων της έννοιας μέσα στη συλλογή και N ο συνολικός αριθμός ατόμων μέσα στη συλλογή.



Σχήμα 23: Παράδειγμα οντολογίας σε OWL για τον τομέα των ταξιδιών. Με διακεκομμένες γραμμές σημειώνονται οι ιδιότητες και σε σκιασμένα ορθογώνια τα στιγμιότυπα των εννοιών.

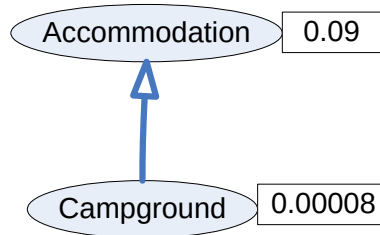
Στον πίνακα που ακολουθεί συγκεντρώνουμε όλες τις περιπτώσεις συσχετίσεων που μας ενδιαφέρουν όπως αυτές προκύπτουν μέσα από το ODM.

Τοποθεσία	A/A	Περιγραφή	Τύπος	Αναπαράσταση
Α. Στο ίδιο δέντρο, κοινή ρίζα C	A1	Ίδιο path Πατέρας-παιδί	ISA	
	A2	Αδέρφια ισοδύναμα	ISA + equivalent	
	A3	Αδέρφια ασύνδετα	ISA + disjoint	
	A4	Αδέρφια επικαλυπτόμενα	siblings	
Β. Σε διαφορετικό δέντρο	B1	Ιδιότητα	ObjectProperty	
	B2	Πεδίο και Εύρος	Domain+Range	
	B3	Αντίστροφη ιδιότητα	InverseOf	

Πίνακας 4: Περιπτώσεις συσχέτισης δύο εννοιών. Στην τρίτη στήλη αποτυπώνονται οι έννοιες σαν σύνολα ανάλογα με τη σχέση που έχουν μεταξύ τους σαν «μάζες» πληροφορίας. Αυτό θα βοηθήσει στον υπολογισμό του κοινού πληροφοριακού περιεχομένου των δύο εννοιών.

Σύμφωνα με τον παραπάνω πίνακα οι διαφορετικές περιπτώσεις συσχετίσεων που διακρίνουμε είναι οι ακόλουθες:

- ❖ A1 Ιεραρχική συσχέτιση (IS-A σχέσεις). Η απλή περίπτωση δύο έννοιες A,B να ανήκουν στην ίδια ιεραρχία και ταυτόχρονα πάνω στο ίδιο μονοπάτι. Πρόκειται για το παράδειγμα δυο οποιονδήποτε εννοιών που συνδέονται απευθείας με τη σχέση της εξειδίκευσης πατέρας-παιδί ή εγγόνι όπως



Σε αυτήν την περίπτωση είτε η ερώτηση είναι συζευκτική (Accommodation AND Campground) είτε διαζευκτική (Accommodation OR Campground) η σχετικότητα της επηρεάζεται μόνο από το περιεχόμενο της πιο ειδικής έννοιας και για τον υπολογισμό της θα έχω ότι

$$sr(Accommodation, Campground) = -\log P(Campground)$$

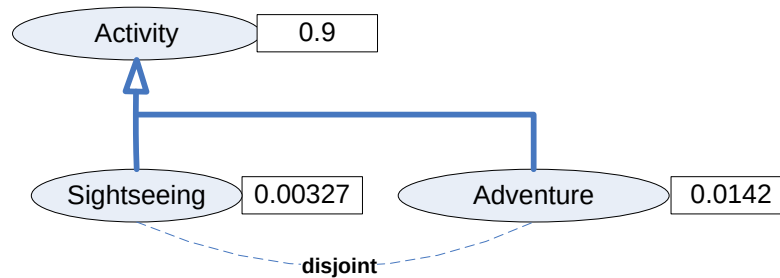
- ❖ A2 Συσχέτιση Ισοδυναμίας μέσα στην ιεραρχία. Η περίπτωση της ισοδυναμίας μεταξύ των εννοιών που είναι κομμάτι της ιεραρχίας. Οι έννοιες θα πρέπει να συνδέονται με την σχέση της ισοδυναμίας ανεξάρτητα από το αν είναι ή όχι μέλη της ίδιας ιεραρχίας. Το παράδειγμα θα μπορούσε να είναι το ακόλουθο:



Διαισθητικά αφού οι έννοιες είναι ισοδύναμες το κοινό τους περιεχόμενο είναι όσο της καθεμίας από τις δύο. Άλλωστε σύμφωνα και με τον ορισμό του ODM δύο ισοδύναμες έννοιες θα έχουν και ακριβώς το ίδιο σύνολο από στιγμιότυπα και επομένως για τον υπολογισμό της σχετικότητάς τους θα έχω τόσο για την περίπτωση της ερώτησης FamilyDestination AND Destination όσο και για την ερώτηση FamilyDestination OR Destination

$$sr(FamilyDestination, Destination) = -\log P(FamilyDestination) = -\log P(Destination)$$

- ❖ A3 Συσχέτιση διαχωρισμένων εννοιών μέσα στην ιεραρχία. Η περίπτωση της αποσύνδεσης δύο εννοιών που ανήκουν στην ίδια ιεραρχία. Εδώ έχουμε τη σχέση της αποσύνδεσης(disjoint) όπως στο ακόλουθο παράδειγμα



Εδώ πέρα όπως είδαμε και στην παράγραφο 4.1.2 οι δύο κλάσεις θα πρέπει να μην έχουν κανένα κοινό στοιχείο στα σύνολα με τα στιγμιότυπά τους. Έτσι για τον υπολογισμό της επιλεκτικότητας θα ισχύει:

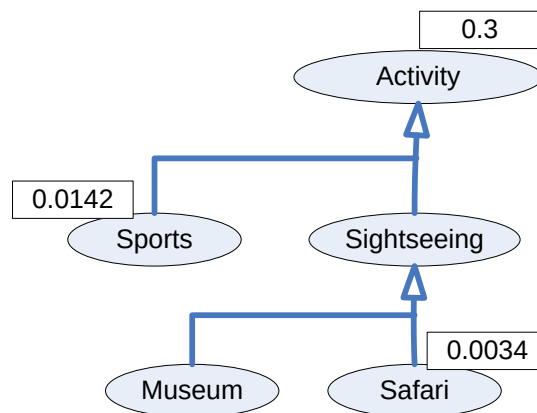
α) για την ερώτηση Sightseeing AND Adventure

$$sr(\text{Sightseeing}, \text{Adventure})=0$$

β) για την ερώτηση Sightseeing OR Adventure

$$sr(\text{Sightseeing}, \text{Adventure}) = -\log P(\text{Sightseeing}) - \log P(\text{Adventure})$$

- ❖ A4 Ασυσχετίστες έννοιες μέσα στην ίδια ιεραρχία. Σε αυτήν την περίπτωση οι έννοιες συσχετίζονται μεταξύ τους μέσω της συγγένειας που έχουν ως μέλη της ίδιας ιεραρχίας και των κοινών στοιχείων που έχουν αποκτήσει από τον κοινό τους πατέρα. Με αυτόν τον τρόπο οι έννοιες έχουν κάποια επικάλυψη μεταξύ τους και μοιράζονται κάποιο κοινό πληροφοριακό περιεχόμενο.



Για τον υπολογισμό της επιλεκτικότητας θα πρέπει να υπολογίσουμε ακριβώς αυτό το κοινό πληροφοριακό περιεχόμενο. Σύμφωνα με όσα είδαμε στην παράγραφο 4.3.1 και την εξίσωση 2 θα έχουμε ότι

α) για την περίπτωση της ερώτησης Sports AND Safari θα ισχύει

$$sr(Sports, Safari) = \frac{2 \log P(Activity)}{\log P(Sports) + \log P(Safari)}$$

β) για την περίπτωση της ερώτησης Sports OR Safari

$sr(Sports, Safari)$

$$= -\log P(Sports) - \log P(Safari) + \frac{2 \log P(Activity)}{\log P(Sports) + \log P(Safari)}$$

- ❖ B1 Συσχέτιση ιδιοτήτων. Η περίπτωση της συσχέτισης δύο εννοιών μέσω μιας ιδιότητας(Property) όπου οι έννοιες A, B έχουν τη θέση του πεδίου (Domain-D) και του εύρους αντίστοιχα(Range-R). Στην περίπτωση αυτή(B1) θα εξετάσουμε την ερώτηση που απευθύνεται σε δύο έννοιες που συνδέονται με μια ιδιότητα και αυτή είναι μέρος της ερώτησης.



Δηλαδή η ερώτηση θα ήταν για has Accommodation με πεδίο τα LuxuryHotel και εύρος ένα UrbanArea. Η ύπαρξη της ιδιότητας είναι απαίτηση του χρήστη. Η σύνδεση αυτή ανεξάρτητα από τις επιλεκτικότητες των όρων που συμμετέχουν στην ερώτηση, περιορίζει το αποτέλεσμα μόνο σε αυτές που συμμετέχουν στη συγκεκριμένη ιδιότητα κάτι που μας οδηγεί στον υπολογισμό της επιλεκτικότητας της ιδιότητας.

Σε αυτό το σημείο εισάγουμε μια καινούρια τιμή που είναι αυτή της επιλεκτικότητας μιας ιδιότητας. Όπως έχουμε δει και κατά την περιγραφή της OWL, οι ιδιότητες ορίζονται με τη βοήθεια ενός χωριστού αντικειμένου αυτού του Property. Επίσης κάθε Property σχετίζεται με ένα πεδίο και ένα εύρος. Ως επιλεκτικότητα ενός Property_x (S_p) ορίζουμε το λόγο

$$S_p(D, R) = \frac{\#Property_x(D, R)}{\#Property_x} \quad \text{Εξ. 8}$$

όπου στον αριθμητή το πλήθος των ιδιοτήτων του συγκεκριμένου τύπου(x) με πεδίο και εύρος τους συγκεκριμένους τύπους που εμφανίζονται στην

ερώτηση και παρονομαστή το πλήθος των ιδιοτήτων του συγκεκριμένου τύπου που εμφανίζονται στην συλλογή ανεξάρτητα από το πεδίο και το εύρος. Στην παρακάτω εικόνα υπάρχει το παράδειγμα της ιδιότητας hasAccommodation. Με βάση το αρχικό παράδειγμα του σχήματος 20 (Σχήμα 23) το hasAccommodation είναι μια ιδιότητα που ορίζεται με εύρος την έννοια Accommodation και πεδίο την έννοια Destination, γενικά. Στο συγκεκριμένο παράδειγμα έχουμε την χρήση της ιδιότητας για δυο συγκεκριμένες υποκλάσεις τις UrbanArea και LuxuryHotel. Επομένως μας ενδιαφέρει να γνωρίζουμε η ιδιότητα hasAccommodation τι επιλεκτικότητα εμφανίζει όχι ανάμεσα σε όλες τις ιδιότητες αλλά συγκεκριμένα για τις ιδιότητες του ίδιου τύπου αφού υπάρχει ελευθερία για το εύρος και το πεδίο που μπορεί να εμφανίζουν.

Συγκεκριμένα, η έννοια UrbanArea είναι στη θέση του πεδίου και η έννοια LuxuryHotel είναι στη θέση του εύρους για την ιδιότητα hasAccommodation. Για την επιλεκτικότητα της συγκεκριμένης ιδιότητας μέσα σε όλη τη συλλογή μπορεί η ιδιότητα hasAccommodation να εμφανίζεται 10 φορές. Αντίστοιχα το hasAccommodation για LuxuryHotel σε UrbanArea μπορεί να εμφανίζεται 3 φορές. Δηλαδή το πληροφοριακό περιεχόμενο της ιδιότητας θα είναι : $IC(\text{hasAccommodation}) = -\log 3/10 = 0.5$.

- ❖ B2 Τυχαία σύνδεση Πεδίου με Εύρους που συνδέονται με κάποιες ιδιότητες.
Σε αυτήν την περίπτωση αναφερόμαστε σε ερώτηση που περιλαμβάνει δύο οποιεσδήποτε έννοιες μέσα στη συλλογή, ανεξαρτήτως της ιεραρχίας και χωρίς να είναι γνωστές άλλες τυχόν συσχετίσεις τους. Δηλαδή, πάνω στο παράδειγμα της περίπτωσης B1 μιλάμε για την ερώτηση UrbanArea AND LuxuryHotel χωρίς να γνωρίζει ο χρήστης και χωρίς να συμπεριληφθεί στην ερώτηση κάποια η ιδιότητα. Εδώ θα πρέπει να λάβουμε υπόψη πως στη γενική περίπτωση δύο έννοιες μπορεί να παρουσιάζουν συσχέτιση με παραπάνω από μια ιδιότητες και άρα να έχει περισσότερες πιθανότητες εμφάνισης ο συνδυασμός τους. Αν ανατρέξουμε στην παράγραφο 0 η πιθανότητα εμφάνισης του συγκεκριμένου συνδυασμού προκύπτει απευθείας από την εξίσωση 5 και είναι ίση με τη πιθανότητα εμφάνισης

όλων των ιδιοτήτων που εμφανίζονται μεταξύ των δύο συγκεκριμένων εννοιών δηλαδή

$$sr(\text{UrbanArea}, \text{LuxuryHotel}) = rel_{op1}(C_1, C_2)$$

- ❖ B3 τυχαία σύνδεση μεταξύ εννοιών που συνδέονται ταυτόχρονα με κάποιες ιδιότητες και κάποιες αντίστροφές τους. Η τελευταία αυτή περίπτωση αναφέρεται στο παράδειγμα της περίπτωσης B2 με τη διαφορά ότι επιπρόσθετα μεταξύ των δύο εννοιών υπάρχουν και συσχετίσεις με τη χρήση της αντίστροφής ιδιότητας inverseOf κάτι που όπως επίσης είδαμε στην παράγραφο 0 ενισχύει τη σύνδεση των εννοιών μεταξύ τους. Αντίστοιχα με την προηγούμενη περίπτωση θα έχουμε τη χρήση της εξίσωσης 6 και το τελικό αποτέλεσμα θα είναι της μορφής:

$$sr(\text{UrbanArea}, \text{LuxuryHotel}) = sr_{op}(C_1, C_2)$$

Στις παραπάνω περιπτώσεις των ιδιοτήτων οι διαζευκτικές ερωτήσεις της μορφής UrbanArea OR LuxuryHotel δεν επηρεάζονται από την ύπαρξη της ιδιότητας και επομένως θα ισχύει για αυτές κανονικά ότι $P(\text{UrbanArea}, \text{LuxuryHotel}) = P(\text{UrbanArea}) + P(\text{LuxuryHotel})$.

4.4. Εφαρμογές

Στο παρόν κεφάλαιο ασχοληθήκαμε με την εκτίμηση της επιλεκτικότητας για ODM δεδομένα και στη συνέχεια με την εκτίμηση της συσχέτισης μεταξύ εννοιών κάτω από το ίδιο πλαίσιο. Αφορμή για τα παραπάνω στάθηκε κατά κύριο λόγο το σύστημα επεξεργασίας και προώθησης ερωτήσεων που αναπτύχθηκε στην παρούσα εργασία. Στόχος, ήταν η διερεύνηση των δυνατοτήτων βελτιστοποίησης των παραπάνω διαδικασιών.

Η εκτίμηση της επιλεκτικότητας αποτελεί το βασικό μέτρο για την εκτίμηση του κόστους των ερωτήσεων και μπορεί να αποτελέσει το βασικό κριτήριο τόσο για την εκτέλεσή τους όσο και για την προώθησή τους. Ειδικά στο κομμάτι της προώθησης όπως είχαμε δει και στη σχετική βιβλιογραφία (Crespo & Garcia-Molina, Routing Indices for Peer-to-Peer Systems, 2002) κ.α.) η επιλεκτικότητα αποτελεί επίσης και ένα καλό μέτρο που μπορεί να λειτουργήσει ως ένδειξη για την παρουσία των δεδομένων στους διάφορους κόμβους του δικτύου. Με τον τρόπο αυτό η

επιλεκτικότητα μπορεί να χρησιμοποιηθεί ως κριτήριο για τη στρατηγική προώθησης των ερωτήσεων σε επίπεδο δικτύου.

Από την άλλη πλευρά σε ένα δίκτυο σαν αυτό που αναπτύχθηκε στα πλαίσια της εργασίας ένα άλλο θέμα που καλούμαστε να αντιμετωπίσουμε είναι αυτό της ετερογένειας των δεδομένων. Από την άποψη αυτή είναι απαραίτητη κάποια μεθοδολογία για την εξεύρεση συσχετίσεων μεταξύ εννοιών που ανήκαν σε μια οντολογία. Είναι φυσικό διαφορετικοί κόμβοι να έχουν κληρονομήσει το ίδιο μοντέλο περιγραφής αλλά να έχουν αναπτύξει και κάποια δικά τους τμήματα είτε ακόμα και να έχουν δημιουργήσει κάποια δικά τους μοντέλα περιγραφής ακολουθώντας όμως το μοντέλου του MOF. Για το λόγο αυτό έπρεπε να είμαστε σε θέση να υπολογίσουμε συσχετίσεις για συνδυασμούς εννοιών προκειμένου να μπορούν να συλλέγονται όλα τα αποτελέσματα για μια ερώτηση με την αντίστοιχη κατάταξή τους.

Κεφάλαιο 5ο - Ανακεφαλαίωση και Συμπεράσματα

5.1. Εισαγωγή

Στο τελευταίο αυτό κεφάλαιο θα προσπαθήσουμε να συγκεντρώσουμε όλα τα κρίσιμα σημεία της εργασίας και να βγάλουμε χρήσιμα συμπεράσματα που έχουν εφαρμογή πέρα από τα πλαίσια της συγκεκριμένη εργασίας. Την ανακεφαλαίωση θα την χωρίσουμε σε δύο κυρίως μέρη, απομονώνοντας το τεχνικό από το ερευνητικό κομμάτι.

5.2. Σχεδιασμός και υλοποίηση

Το σημαντικότερο στοιχείο για την παρούσα εργασία είναι το γεγονός ότι περιελάμβανε την ανάπτυξη ενός ολοκληρωμένου συστήματος. Σίγουρα κάποια τμήματα του συστήματος αναπτύχθηκαν στο πλαίσιο του ερευνητικού προγράμματος DBE όμως πριν από αυτά, πολύ σημαντικό ρόλο έπαιξε η ανάλυση των απαιτήσεων και ο σχεδιασμός του συστήματος. Έπρεπε να επιλεγούν οι τεχνολογίες που θα χρησιμοποιούνταν για να δημιουργηθεί το τεχνολογικό πλαίσιο ανάπτυξης του συστήματος και ταυτόχρονα να γίνει η σχεδίαση μιας αρχιτεκτονικής που θα ικανοποιούσε όλες τις αρχικές απαιτήσεις αλλά και τη βιωσιμότητα του συστήματος σε πραγματικές συνθήκες και διαχρονικά.

Κοινός άξονας του συστήματος στάθηκε η σημασιολογική περιγραφή των δεδομένων με τον ορισμό του ODM. Πάνω στην περιγραφή στηρίχθηκε τόσο η δημιουργία του υπερκείμενου δικτύου όσο και η ανάπτυξη των μηχανισμών ανάκτησης.

5.2.1. Το δίκτυο

Το πρώτο κομμάτι περιελάμβανε την ανάπτυξη ενός δικτύου μεταξύ των επιχειρήσεων που μετείχαν στο πρόγραμμα. Για το σκοπό αυτό είδαμε πως η κατασκευή ενός ομότιμου δικτύου ήταν αυτή που επέτρεπε την αυτόνομη και δυναμική λειτουργία του δικτύου χωρίς την ανάγκη κεντρικής διαχείρισης ή κόμβων με «ειδικές» αρμοδιότητες. Σε αυτό το σημείο μιλάμε για το φυσικό δίκτυο και για το λογισμικό που χρειάστηκε να διαθέτει κάθε κόμβος ώστε να είναι μέλος του δικτύου και να απολαμβάνει τις υπηρεσίες του. Απαιτήθηκε προσαρμογή στον πυρήνα της ανοικτού-κώδικα πλατφόρμας FADA ώστε να δημιουργηθεί ένα δίκτυο

με τις επιθυμητές ιδιότητες. Το αποτέλεσμα δοκιμάστηκε σε πραγματικές συνθήκες εγκαθιστώντας το λογισμικό, τόσο σε κοντινούς και εσωτερικούς κόμβους όπως τα μηχανήματα του εργαστηρίου Διανεμημένων Πληροφοριακών Συστημάτων και εφαρμογών, όσο και σε απομακρυσμένους και περισσότερους κόμβους που ήταν είτε συνεργάτες του προγράμματος είτε πραγματικές επιχειρήσεις που μετείχαν σε αυτό. Ειδικότερα στα πλαίσια της εργασίας χρειάστηκε να ενσωματωθούν στο συγκεκριμένο λογισμικό όλες οι υπηρεσίες που θα έπρεπε να είναι διαθέσιμες στους διάφορους κόμβους του δικτύου προκειμένου να εξυπηρετηθούν οι ανάγκες για ανάκτηση και διαχείριση της Βάσης Γνώσης.

5.2.2. Η Βάση Γνώσης και το ODM

Η Βάση Γνώσης του συστήματος αποτέλεσε ένα πολύ σημαντικό κομμάτι της εργασίας καθώς έπρεπε από τη μια πλευρά να επιλεγεί ο τρόπος περιγραφής όλων των δεδομένων του συστήματος και από την άλλη πλευρά να υλοποιηθούν οι μηχανισμοί για τη διαχείριση και ανάκτησή τους. Σε αυτό το σημείο έγινε χρήση οντολογιών που αναπτύχθηκαν για να καλύψουν τα διαφορετικά είδη δεδομένων και μεταδεδομένων του συστήματος εξασφαλίζοντας:

1. την ελευθερία στον τελικό χρήστη για την περιγραφή του με εξατομικευμένο τρόπο (προσαρμοσμένη οντολογία)
2. την επεκτασιμότητα του συστήματος ενσωματώνοντας μελλοντικούς χρήστες και από διαφορετικές (θεματικές) περιοχές
3. τη διαχείριση των δεδομένων σε διαφορετικά επίπεδα αφαίρεσης (ανεξάρτητη οπτική στα μεταδεδομένα και δεδομένα του συστήματος)
4. την αντιμετώπιση του προβλήματος της ετερογένειας των δεδομένων αφού αυτά μπορούσαν να βασίζονται σε διαφορετικές οντολογίες
5. την αποθήκευση και διαχείριση των δεδομένων σε επίπεδο δικτύου
6. τη δημιουργία και εξυπηρέτηση σύνθετων ερωτήσεων τόσο σε τοπικό επίπεδο όσο και σε επίπεδο δικτύου.

Για όλα τα παραπάνω καθοριστικό ρόλο έπαιξε η αρχιτεκτονική του συστήματος και η χρήση οντολογιών. Πιο συγκεκριμένα όπως είδαμε στο Κεφάλαιο 2 (§2.4), επιλέχτηκε η MDA αρχιτεκτονική ενώ μοντέλα οντολογιών σχεδιάστηκαν από την αρχή ώστε να καλύπτουν όλους τους τύπους δεδομένων στο σύστημα. Στο πλαίσιο αυτό, ορίστηκε και αναπτύχθηκε μια γλώσσα περιγραφής οντολογιών, το ODM, η

οποία αποτέλεσε και τον κεντρικό άξονα του συστήματος για την περιγραφή όλων των δεδομένων και μεταδεδομένων. Όλα τα υπόλοιπα μοντέλα περιγραφής που απαιτήθηκαν στα πλαίσια του προγράμματος κάνανε χρήση του ODM σαν μέσο κοινής περιγραφής του «λεξιλογίου» που χρησιμοποιούσαν. Τεχνικά, στη Βάση Γνώσης, οι οντολογίες αντιμετωπίστηκαν σαν XML έγγραφα και όλο το κομμάτι της διαχείρισης και ανάκτησής τους στηρίχθηκε σε ανάπτυξη των αντίστοιχων υπηρεσιών πάνω από μια XML βάση δεδομένων. Οι υπηρεσίες αυτές ήταν ορατές σε όλα τα μέλη του δικτύου. Επίσης όλοι οι μηχανισμοί ανάκτησης που συνοψίζουμε στη συνέχεια εκμεταλλεύτηκαν τη σημασιολογία και το υπερκείμενο δίκτυο ώστε να επιτύχουν τη μέγιστη απόδοση με βάση το πλαίσιο που είχε οριστεί.

5.2.3. Μηχανισμοί Ανάκτησης

Όπως είχαμε δει και από την αρχή οι μηχανισμοί ανάκτησης των δεδομένων ήταν από τους βασικούς στόχους του συστήματος. Η πρόκληση ήταν διπλή γιατί έπρεπε από τη μια πλευρά να λαμβάνουν υπόψη την ετερογένεια των δεδομένων και από την άλλη πλευρά να έχουν εφαρμογή στο ομότιμο δίκτυο. Για το σκοπό αυτό δημιουργήθηκαν γενικευμένοι μηχανισμοί ανάκτησης από τις οντολογίες λαμβάνοντας υπόψη και την ιεραρχία των δεδομένων με τα διαφορετικά επίπεδα αφαίρεσης. Κεντρικό ρόλο έπαιξε η περιγραφή των δεδομένων με το ODM στη Βάση Γνώσης. Ο σχεδιασμός και η αρχιτεκτονική του συστήματος επέτρεψαν τη δημιουργία ενός ενιαίου μηχανισμού ανάκτησης ανεξάρτητα από τις ανάγκες και τα δεδομένα κάθε κόμβου.

Συνοψίζοντας τους στόχους που επιτεύχθηκαν στο κομμάτι αυτό, έχουμε:

- ❖ την ανάπτυξη υπηρεσιών ανάκτησης ενσωματωμένων στον πυρήνα λογισμικού κάθε κόμβου του δικτύου
- ❖ την αναγνώριση και ανάλυση σημασιολογικά πλούσιων ερωτήσεων διατυπωμένων με τη βοήθεια της QML
- ❖ την ανάπτυξη γενικευμένου μηχανισμού παραγωγής ερωτήσεων για την τελική εκτέλεσή τους σε διαφορετικές πλατφόρμες αποθήκευσης
- ❖ τη δυνατότητα δημιουργίας επεκταμένων ερωτήσεων για την κάλυψη των περιπτώσεων αντιστοίχισης μεταξύ διαφορετικών οντολογιών
- ❖ τη δημιουργία υπηρεσίας συστάσεων με βάση προφίλ χρηστών του συστήματος

❖ τη δημιουργία μηχανισμών συλλογής των αποτελεσμάτων από το δίκτυο

Πιο αναλυτικά, αναπτύχθηκε μηχανισμός επέκτασης των ερωτήσεων για να καλύψει όλες τις διαφορετικές μεταξύ τους οντολογίες. Σημαντικό ρόλο και κομμάτι της εργασίας έπαιξε η Μονάδα Πρόσβασης στη Γνώση (3.5). Αυτή ήταν υπεύθυνη για την ανάλυση των ερωτήσεων και την επεξεργασία τόσο για την επέκτασή τους όσο και για την προώθησή τους. Δεν θα πρέπει να παραλείψουμε τη μετατροπή των ερωτήσεων σε ασαφείς ώστε να καλυφτούν και πιο σύνθετες ερωτήσεις των χρηστών ή ερωτήσεις που να απαιτούν κάποια κατάταξη των αποτελεσμάτων. Φυσικά όλα τα παραπάνω αναπτύχθηκαν στο πλαίσιο του δικτύου παρέχοντας τη δυνατότητα συλλογής αποτελεσμάτων από όλους τους διαθέσιμους κόμβους κάθε στιγμή. Παράλληλα, το σύστημα παρείχε τη δυνατότητα στο χρήστη για τη δημιουργία ερωτήσεων σε επίπεδο οντολογιών (όχι μόνο δεδομένων) ενώ ο μηχανισμός ανάλυσης των ερωτήσεων διέθετε ένα γενικευμένο μηχανισμό παραγωγής των τελικών ερωτήσεων προς την τελική βάση δεδομένων και στην απαιτούμενη γλώσσα ερωτήσεων. Χαρακτηριστικά, στα πλαίσια της εργασίας είδαμε την μετατροπή του μηχανισμού για την παραγωγή ερωτήσεων από σχεσιακό μοντέλο σε ημιδομημένο μοντέλο περιγραφής των δεδομένων.

5.2.4. Κόστος Ερωτήσεων

Το κομμάτι της ανάλυσης των ερωτήσεων και της προώθησής τους σε ένα ομότιμο δίκτυο αποτελεί σίγουρα μια από τις σύγχρονες ερευνητικές προκλήσεις. Η αποδοτική προώθηση στο δίκτυο έχει αποτελέσει αντικείμενο μεγάλου μέρους της βιβλιογραφίας στην περιοχή, χωρίς όμως να υπάρχουν γενικές λύσεις αφού κάθε δίκτυο φέρει και κάποια δικά του χαρακτηριστικά ή έχει κάποιες διαφορετικές απαιτήσεις. Το σημασιολογικό δίκτυο που αναπτύχθηκε στην παρούσα εργασία συγκέντρωνε πολλές από τις σύγχρονες τάσεις, έχοντας αρκετά πρωτότυπα χαρακτηριστικά σαν ολοκληρωμένο σύστημα. Στο πλαίσιο αυτό ασχοληθήκαμε με το θέμα της ανάλυσης του κόστους των ερωτήσεων πάνω από οντολογίες. Είδαμε πως παρ' όλο που το θέμα του κόστους των ερωτήσεων είναι παλιό και διαδεδομένο τόσο στα παραδοσιακά συστήματα βάσεων δεδομένων όσο και σε σχέση με XML έγγραφα, δεν υπήρχε κάποια συγκεκριμένη εργασία για το θέμα της ανάλυσης του κόστους ερωτήσεων πάνω από OWL οντολογίες. Με αφορμή αυτό ασχοληθήκαμε με τη προσαρμογή και μεταφορά κάποιον γνωστών μεθοδολογιών

και εφαρμογών, προτείνοντας μια μεθοδολογία για την εκτίμηση του κόστους των ερωτήσεων. Η μεθοδολογίες που προτάθηκαν βασίζονται σε δύο κατευθύνσεις:

- τη σύνθεση γνωστών τιμών επιλεκτικότητας με τη βοήθεια της υπόθεσης της ανεξαρτησίας
- την προσθήκη στις τιμές που μπορεί να έχει προϋπολογισμένες μια αρμόδια υπηρεσίας καινούριων στατιστικών μετρικών όπως η επιλεκτικότητα μιας ιδιότητας

Η τιμή της επιλεκτικότητας είναι χρήσιμη σε μια σειρά εφαρμογών. Κατ' αρχήν από τα παραδοσιακά συστήματα ακόμα, είναι χρήσιμη για τον υπολογισμό του κόστους των ερωτήσεων ενώ ανεξάρτητα από την πλατφόρμα εργασίας και για τη λήψη αποφάσεων για τη βελτιστοποίηση της εκτέλεσής της. Επιπρόσθετα σε ένα πλαίσιο σαν αυτό που αναπτύξαμε, δίνει τη δυνατότητα λήψης αποφάσεων δρομολόγησης των ερωτήσεων σε επίπεδο δικτύου με βάση τη σημασιολογία τους. Το τελευταίο είναι πολύ σημαντικό για τη βελτίωση της απόδοσης των μηχανισμών ανάκτησης σε περιπτώσεις ομότιμων δικτύων.

Στη συνέχεια του κεφαλαίου 4 συγκεντρώσαμε τις συσχετίσεις που εμφανίζονται σε δεδομένα οντολογιών μέσω των γνωστών σχέσεων και αξιωμάτων με βάση το ODM (ή/και την OWL) και προτείναμε μια σειρά αντίστοιχων μετρικών για τον υπολογισμό της σημασιολογικής σχετικότητας δύο εννοιών. Αυτό που αποτέλεσε κίνητρο για τη δουλειά και είναι ίσως και το πιο σημαντικό κομμάτι για μελλοντική διερεύνηση ήταν η δυνατότητα συνδυασμού της συσχέτισης των εννοιών με την επιλεκτικότητα ώστε να μπορέσουμε τελικά να οδηγηθούμε στην απόρριψη της υπόθεσης της ανεξαρτησίας.

Βιβλιογραφία

- B. Omelayenko, M. K. (Επιμ.). (2002). Proceedings. *Workshop on Knowledge Transformation for the Semantic Web (KTSW)*. Lyon, France: www.cs.vu.nl/~borys/events/KTSW02.
- Aberer, K. (2001). P-grid: a self-organizing access structure for p2p information systems. *Sixth international conference on cooperative information systems (CoopIS 2001)*. Trento, Italy: Springer Verlag, <http://www.p-grid.org/Papers/CoopIS2001.pdf>.
- Aberer, K., Cudr'e-Mauroux, P., Hauswirth, M., & Pelt, V. (2004). Gridvine: building internetscale semantic overlay networks. *3rd international semantic web conference (ISWC2004)* (σσ. 107–121). Hiroshima, Japan: Springer.
- Abidi, S., & Pang, X. (2004). Knowledge sharing over P2P knowledge networks: A peer ontology and semantic overlay driven approach. *International Conference on Knowledge Management*.
- Abounaga, A., Alameldeen, A. R., & Naughton, J. F. (2001). Estimating the Selectivity of XML Path Expressions for Internet Scale Applications. *27th VLDB*. Roma, Italy.
- Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, Volume 16, p 3-9.
- Aduna Open Source. (n.d.). *OpenRDF.org*. Ανάκτηση 12 2010, από <http://www.openrdf.org/>
- Al-Mubaid, H., & Nguyen, H. A. (2007). Semantic Distance of Concepts within a Unified Framework in the Biomedical Domain. *SAC*. Seoul, Korea: ACM.
- Arenas, M., Kantere, V., Kementsietsidis, A., Kiringa, I., Miller, R. J., & Mylopoulos, J. (2003). The Hyperion Project: From Data Integration to Data Coordination. *SIGMOD Record, Special Issue on Peer-to-Peer Data Management*, (σσ. 53-58).
- Atwood, T., & Cattell, R. G. (1994). *The Object database standard /ODMG-93*. San Mateo: Morgan-Kaufmann.
- Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., και συν. (2004). *OWL Web Ontology Language Reference*. Ανάκτηση 2004, από <http://www.w3.org/TR/owl-ref/>
- Bellinger, G. (2004). *Knowledge management: emerging perspectives*. Ανάκτηση 5 10, 2009, από <http://www.systems-thinking.org/kmgmt/kmgmt.htm>
- Bergeron, B. (2003). *Essentials of Knowledge Management*. Hoboken, New Jersey: John Wiley & Sons, Inc.

- Berners-Lee, T. (1998). *Semantic Web Road Map*. Ανάκτηση 2004, από W3C Design Issues: <http://www.w3.org/DesignIssues/Semantic.html>
- Bertino, E. (1994). Index Configuration in Object-Oriented Databases. *VLDB Journal*, 355-399.
- Bohrer, W., Brice, S. R., Cichocki, A., Fowler, J., Helal, A., Kashyap, V., και συν. (1997). Infosleuth: semantic integration of information in open and dynamic environments. *ACM SIGMOD international conference on management of data* (σσ. 195–206). Tucson, Arizona, USA: ACM Press.
- Booch, G., Rumbaugh, J., & Jacobson, I. (1998). *The Unified ModelingLanguage User Guide*. Massachusetts: Addison-Wesley.
- Bray, T., Paoli, J., Sperberg-McQueen, C. M., & Maler, E. (2000). *Extensible Markup Language (XML) 1.0 (Second Edition)*. Ανάκτηση 2002, από W3C Recommendation: <http://www.w3.org/TR/2000/REC-xml-20001006>
- Brickley, D., & Guha, V. R. (2000). *Resource Description Framework (RDF) Schema Specification 1.0*. Ανάκτηση 2004, από W3C: <http://www.w3.org/TR/2000/CR-rdf-schema-20000327>
- Brunkhorst, I., Dhraief, H., Kemper, A., Nejd, W., & Wiesner, C. (2003). Distributed Queries and Query Optimization in Schema-Based P2P-Systems. *VLDB*.
- Bruno, N., & Chaudhuri, S. (2004). Conditional Selectivity for Statistics on Query Expressions. *SIGMOD*. Paris, France: ACM.
- Byers, J., Considine, J., & Mitzenmacher, M. (2001). Simple load balancing for distributed hash tables. *International workshop on design issues in anonymity and unobservability*, (σσ. 46–66).
- Castano, A. (2003). Ontology-Addressable Contents in P2P Networks. *1st Workshop on Semantics in Peer-to-Peer and Grid Computing*. Budapest.
- Castano, S., De Antonellis, V., & De Capitani, S. (2001). Global Viewing of Heterogeneous Data Sources. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 13(2), 277-297.
- Castano, S., Ferrara, A., & Montanelli, S. (2003). H-match: an Algorithm for Dynamically Matching Ontologies in Peer-based Systems. *1st SWDB VLDB Workshop*. Berlin, Germany.
- Castano, S., Ferrara, A., Montanelli, S., Pagani, E., & Rossi, G. P. (2003). *WWW'03 1st SemPGRIDWorkshop*, (σσ. Ontology-Addressable Contents in P2P networks).
- Chan, C.-Y., Felber, P., Garofalakis, M., & Rastogi, R. (2002). Efficient filtering of XML documents with XPath expressions. *VLDB Journal*, 354-379.

- Cho, W.-S., Hong, K.-H., & Loh, W.-K. (2007). Estimating nested selectivity in object-oriented and object-relational databases. *Information and Software Technology*, 806-816.
- Christodoulakis, S. (1983). Estimating Record Selectivities. *Information Systems*, 8(2), 105-115.
- Christodoulakis, S. (1984, June). Implications of Certain Assumptions in Database Performance Evaluation. *ACM Transactions on Database Systems*, 9(2), 163-186.
- Clarke, I., Sandberg, O., Wiley, B., & Hong, T. (2000). Freenet: A distributed anonymous information storage and retrieval system. *ICSI Workshop on Design Issues in Anonymity and Unobservability*. Berkeley, California: freenet.sourceforge.net.
- Clarke, I., Sandberg, O., Wiley, B., & Hong, T. (June 2000). Freenet: A distributed anonymous information storage and retrieval system. *Proceedings of ICSI Workshop on Design Issues in Anonymity and Unobservability*. Berkeley, California: freenet.sourceforge.net.
- Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley-Interscience.
- Crainiceanu, A., Linga, P., Gehrke, J., & Shanmugasundaram, J. (2004). Querying Peer-to-Peer Networks Using PTrees. *Seventh International Workshop on the Web and Databases (WebDB)*. Paris, France.
- Crespo, A., & Garcia-Molina, H. (2002). Routing indices for peer-to-peer systems. *22nd international conference on distributed computing systems (ICDCS'02)* (σ. 23). Washington, DC, USA: IEEE Computer Society.
- Crespo, A., & Garcia-Molina, H. (2002). Routing Indices for Peer-to-Peer Systems. *22nd International Conference on Distributed Computing Systems (ICDCS'02)*.
- Crespo, A., & Garcia-Molina, H. (2004). Semantic Overlay Networks for P2P Systems. *AP2PC*, (σσ. 1-13).
- Diao, Y., Altinel, M., Franklin, M. J., Zhang, H., & Fischer, P. (2003, December). Path Sharing and Predicate Evaluation for High-Performance XML Filtering. *ACM Transactions on Database Systems*, 28(4), 467-516.
- Doulkeridis, C., Vlachou, A., Kotidis, Y., & Vazirgiannis, M. (2007). Peer-to-Peer Similarity Search in Metric Spaces. *VLDB*. Vienna, Austria: VLDB Endowment, ACM.
- Elenius, D., & Ingmarsson, M. (2004). Ontology-based Service Discovery in P2P Networks. *1st international workshop in P2P Knowledge Management, P2PKM*.
- Finin, T., Fritzson, R., McKay, D., & McEntire, R. (1994). KQML as an Agent Communication Language. Στο B. B. Adam N (Επιμ.), *3rd international conference on information and knowledge management (CIKM'94)* (σσ. 456-463). Gaithersburg, MD, USA: ACM Press.

- Formica, A. (2007, July). Similarity of XML-Schema Elements: A Structural and Information Content Approach. *The Computer Journal*.
- Garcia-Molina, H., & Yang, B. (2002). Efficient Search in Peer-to-peer Networks. *ICDCS*.
- Gardarin, G., Gruser, J.-R., & Tang, Z.-H. (1996). Cost-based Selection of Path Expression Processing Algorithms in Object-Oriented Databases. *22nd VLDB* (σσ. 390-401). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Getoor, L., Taskar, B., & Koller, D. (2001). Selectivity Estimation using Probabilistic Models. *ACM SIGMOD*. ACM.
- GNU. (n.d.). *Regarding Gnutella*. Ανάκτηση 2009, από GNU Operating System: <http://www.gnu.org/philosophy/gnutella.html>
- Graefe, G. (1993). Query Evaluation Techniques for Large Databases. *ACM Computing Surveys*, 25(2), 75-170.
- Gravano, L., Garcia-Molina, H., & Tomasic, A. (1994?). *Precision and Recall of GLOSS Estimators for Database Discovery*. Stanford.
- Gravano, L., Garcia-Molina, H., & Tomasic, A. (1994?). *The Effectiveness of Gloss for the Text Database Discovery Problem*. Stanford.
- Gruber, T. G. (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Στο N. Guarino, & R. Poli, *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Kluwer Academic Publishers.
- Gruber, T. R. (2003). *What is an Ontology?* Retrieved 5 6, 2009, from <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>
- Guo, Y., Pan, Z., & Heflin, J. (2005). LUBM: A Benchmark for OWL Knowledge Base Systems. *Journal of Web Semantics*, 3(2), 158-182.
- Haase, P., Broekstra, J., Ehrig, M., Menken, M., Mika, P., Plechawski, M., και συν. (n.d.). Bibster—a semantics-based bibliographic peer-to-peer system. *3rd international semantic web conference*. Hiroshima, Japan: <http://www.aifb.uni-karlsruhe.de/WBS/pha/publications/bibster04iswc.pdf>.
- Haase, P., Siebes, R., & van Harmelen, F. (2004). Peer selection in peer-to-peer networks with semantic topologies. Στο M. Bouzeghoub (Επιμ.), *International conference on semantics in a networked world (ICNSW'04)*. 3226, σσ. 108–125. Paris: Springer Verlag.
- Haase, P., Siebes, R., & van Harmelen, F. (2007). Expertise-based peer selection in Peer-to-Peer networks. *Knowledge Information Systems*. Springer-Verlag London Limited.

- Halevy, Y. A., Ives, G. Z., Mork, P., & Tatarinov, I. (May 2003). Piazza: data management infrastructure for semantic web applications. *Twelfth International World Wide Web Conference, WWW2003* (σσ. 556–567). Budapest, Hungary: ACM.
- Hariri, B. B., Abolhassani, H., & Khodaei, A. (2006). A new Structural Similarity Measure for Ontology Alignment. *SWWS*, (σσ. 36-42).
- Hose, K., Karnstedt, M., Sattler, K.-U., & Stehr, E.-A. (2005). Adaptive Routing Filters for Robust Query Processing in Schema-Based P2P Systems. *9th International Database Engineering & Application Symposium (IDEAS'05)*. IEEE.
- Java Metadata Interface Specification (v1.0)*. (n.d.). Ανάκτηση από <http://java.sun.com/products/jmi/index.jsp>
- Jiang, J. J., & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *International Conference Research on Computational Linguistics(ROCLING X)*. Taiwan.
- Joseph, S. (2002). NeuroGrid: Semantically Routing Queries in Peer-to-Peer Networks. *Intl. Workshop on Peer-to-Peer Computing*.
- Joseph, S., & Hoshiai, T. (2003, June). Decentralized Meta-Data Strategies: Effective Peer-to-Peer Search. *IEICE Transactions Communication, E86-B(6)*, 1740-1753.
- Kalogeraki, V., Gunopulos, D., & Zeinalipour-Yaz, D. (2002). A Local Search Mechanism for Peer-to-Peer Networks. *Eleventh International Conference on Information and Knowledge Management (CIKM)*. McLean.
- Kantere, V., Kiringa, I., Mylopoulos, J., Kementsietsidis, A., & Arenas, M. (September 2003). Coordinating Peer Databases Using ECA Rules. *International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P)*.
- Kotopoulos, G., Kazasis, F., & Christodoulakis, S. (2007). Querying MOF Repositories: The Design and Implementation of the Query Metamodel Language (QML). *Digital EcoSystems and Technologies Conference, DEST '07. Inaugural IEEE-IES*, (σσ. 373-378).
- Leibowitz, N., Ripeanu, M., & Wierzbicki, A. (2003). Deconstructing the kazaa network. *3rd IEEE workshop on internet applications (WIAPP'03)*. Santa Clara, CA.
- Li, J., & Vuong, S. (2005). Ontology-Based Clustering and Routing in Peer-to-Peer Networks. *Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT05)*. IEEE.
- Li, M., Lee, W., Sivasubramaniam, A., & Lee, D. L. (2004). A Small World Overlay Network for Semantic Based Search in P2P Systems. *Workshop on Semantics in Peerto-Peer and Grid Computing (SemPGrid)*.

- Lim, L., Wang, M., Padmanabhan, S., Vitter, J. S., & Parr, R. (2002). XPathLearner: An On-Line Self-Tuning Markov Histogram for XML Path Selectivity Estimation. *28th VLDB*. Hong Kong, China.
- Lin, D. (1998). An Information-Theoretic Definition of Similarity. *15th International conference on Machine Learning*.
- Löser, A. (2004). Towards Taxonomy based Routing in P2P Networks. *Second Workshop on Semantics in Peer-to-Peer and Grid Computing at the Thirteenth International World Wide Web Conference*.
- Löser, A., Staab, S., & Tempich, C. (2005). Semantic Methods for P2P Query Routing. *MATES*.
- Löser, A., Staab, S., & Tempich, C. (2007, January). Semantic Social Overlay Networks. *IEEE Journal on Selected Areas in Communications*, 25(1).
- Löser, A., Wolpers, M., Siberski, W., & Nejd, W. (2003). Efficient data store discovery in a scientific P2P network. *CEUR WS 83. Colocated with the 2. ISWC-03*. Ashish N, Goble C.
- Luo, C., Jiang, Z., Hou, W.-C., Yan, F., & Zhu, Q. (2008). A relational model for XML structural joins and their size estimations. *Knowledge Information Systems* (σσ. 97-127). Springer-Verlag London Limited.
- Lv, Q., Cao, P., Cohen, E., Li, K., & Shenker, S. (2002). Search and replication in unstructured peer-to-peer networks. *International Conference on Supercomputing (ICS'02)*. ACM.
- Maedche, A., Motik, B., Stoljanovic, L., Studer, R., & Volz, R. (2003). An infrastructure for searching, reusing and evolving distributed ontologies. *WWW2003*. Budapest.
- McIlraith, S., Son, T., & Zeng, H. (2001, March/April). Semantic web services. *IEEE intelligent systems*(Special Issue on the Semantic Web),
citeseer.ist.psu.edu/mcilraith01semantic.html.
- Menascie, D., & Kanchanapalli, L. (2002, September). Probabilistic Scalable P2P Resource Location Services. *ACM Sigmetrics Performance Evaluation Review*, 30(2).
- Moore, J. (1996). *The death of competition: leadership and strategy in the age of business ecosystems*. New York: Harperbusiness.
- Napster. (n.d.). Ανάκτηση από <http://www.napster.com>
- Nejd, W. (2003). Super-Peer-Based Routing and Clustering Strategies for RDF-Based Peer-To-Peer Networks. *ACM International Conference on World Wide Web, WWW2003*.
- Nejd, W. (May 2002). EDUTELLA: a P2P networking infrastructure based on RDF. *International World Wide Web Conference 2002 (WWW2002)*, (σσ. 604–615). Honolulu, Hawaii, USA.

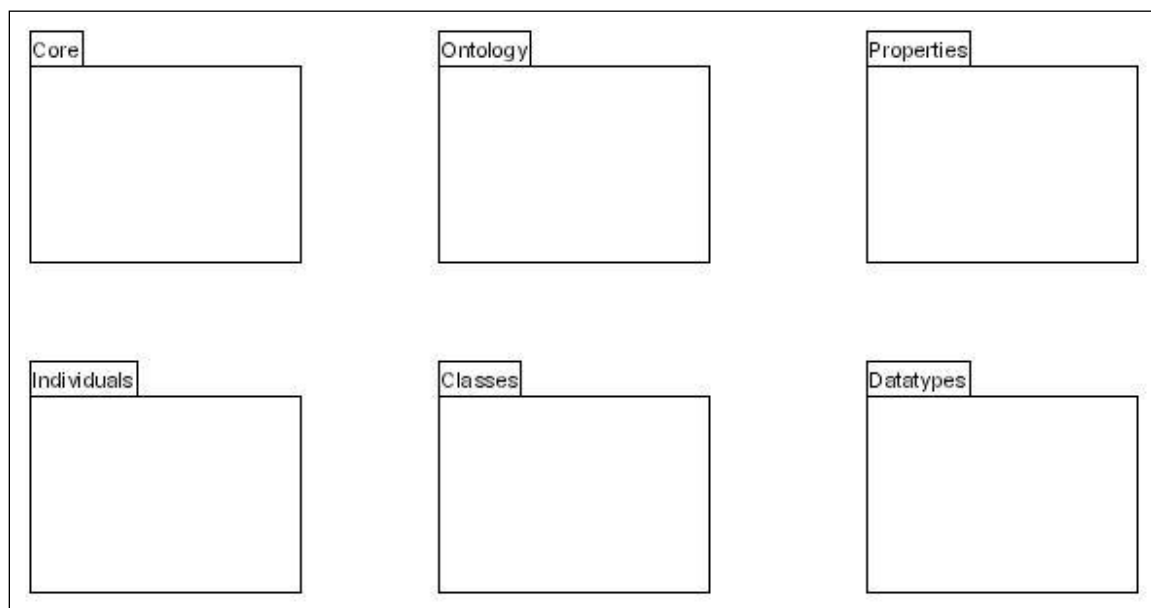
- Nejdl, W., Wolpers, M., Siberski, W., Schmitz, C., Schlosser, M., Brunkhorst, I., και συν. (2003). Super-peer-based routing and clustering strategies for rdf-based peer-to-peer networks. *12th international world wide web conference*. Budapest.
- NetBeans MetaData Repository. (n.d.). Ανάκτηση από <http://mdr.netbeans.org/>
- Ng, W. S., Ooi, B. C., Tan, K.-L., & Zhou, A. Y. (2003). *19th International Conference on Data Engineering, ICDE*.
- Nonaka, I., & Takeuchi, H. (1995). *The Knowledge Creating Company*. Oxford University Press.
- OMG. (2002). *OMG XMI Specification, v1.2*. Ανάκτηση 2003, από Object Management Group: <http://www.omg.org/cgi-bin/doc?formal/2002-01-01>
- OMG. (n.d.). *MDA*. Ανάκτηση 2004, από OMG Model Driven Architecture : <http://www.omg.org/mda/>
- OMG. (n.d.). *The Object Management Group, ODM RFP*. Ανάκτηση από <http://www.omg.org/cgi-bin/apps/doc?ad/03-09-06.pdf>
- Oram, A. (2001). Gnutella. Στο *Peer-to-peer: harnessing the power of disruptive technologies* (σσ. 94–122). O'Reilly and Associates.
- Piatetsky-Shapiro, G., & Connell, C. (1984). Accurate estimation of the number of tuples satisfying a condition. *ACM*.
- Polizotis, N., & Garofalakis, M. (2006, September). XSKETCH Synopses for XML Data Graphs. *ACM Transactions on Database Systems*, 31(3), 1014-1063.
- Polyzotis, N., & Garofalakis, M. (2002). Statistical Synopses for GraphStructured XML Databases. *ACM SIGMOD*. Madison, Wisconsin, USA: ACM.
- Polyzotis, N., & Garofalakis, M. (2002). Structure and Value Synopses for XML Data Graphs. *28th VLDB*. Hong Kong, China.
- Polyzotis, N., Garofalakis, M., & Ioannidis, Y. (2004). Selectivity Estimation for XML Twigs. *20th International Conference on Data Engineering (ICDE)*. IEEE Computer Society.
- Poosala, V., & Ioannidis, Y. (1997). Selectivity Estimation Without the Attribute Value Independence Assumption. *23rd VLDB*. Athens, Greece.
- Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, 334–350.
- Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *VLDB Journal*, 334-350.

- Ratnasamy, S., Francis, P., Handley, M., Karp, R., & Shenker, S. (2001). A scalable content-addressable network. *ACM SIGCOMM*.
- Ratnasamy, S., Francis, P., Handley, M., Karp, R., & Shenker, S. (August 2001). A scalable content-addressable network. *ACM SIGCOMM*.
- Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *IJCAI*.
- Resnik, P. (1998). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Artificial Intelligence Research*.
- Richardson, R., & Smeaton, A. F. (1995). *Using WordNet in a Knowledge-Based Approach to Information Retrieval*. Working Paper, Dublin City University, School of Computer Applications, Ireland.
- Ross, S. (1997(1976)). *A first course in probability* (5 εκδ.). California: Prentice Hall.
- Rostami, H., Habibi, J., Abolhassani, H., Amirkhani, M., & Rahnama, A. (2006). An Ontology Based Local Index in P2P Networks. *Second International Conference on Semantics, Knowledge, and Grid(SKG)* (σ. 11). IEEE Computer Society: Washington, DC, USA .
- Rowstron, A., & Druschel, P. (2001). Pastry: Scalable, distributed object location and routing for largescale peer-to-peer systems. *18th IFIP/ACM Int'l. Conf. Distributed Systems Platforms*.
- Rowstron, A., & Druschel, P. (2001). Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. *IFIP/ACM international conference on distributed systems platforms (Middleware)* (σσ. 329–350). Heidelberg, Germany: <http://research.microsoft.com/antr/PAST/pastry.pdf>.
- Ruckhaus, E., Ruiz, E., & Vidal, M. E. (2008). Query evaluation and optimization in the semantic web. *Theory and Practice of Logic Programming*, 8(3), 393-409.
- Sahin, O. D., Gupta, A., Agrawal, D., & El Abbadi, A. (2002). *Query Processing Over Peer-To-Peer Data Sharing Systems*. UCSB Technical Report.
- Sangpachatanaruk, C., & Znati, T. (2004). Semantic Driven Hashing (SDH):An Ontology-based Search Scheme for the Semantic Aware Network (SA Net). *Fourth International Conference on Peer-to-Peer Computing (P2P'04)*. IEEE.
- Sartiani, C., Manghi, P., Ghelli, G., & Conforti, G. (2004). XPeer: A Self-Organizing XML P2P Database System. *EDBT Workshops*, (σσ. 456-465).

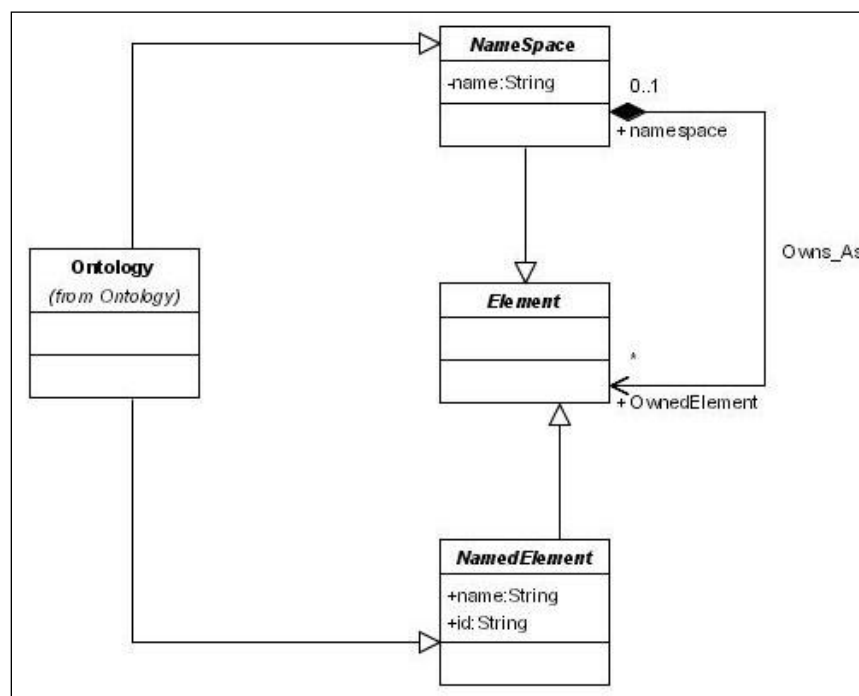
- Sartiani, C., Marciniak, M., & Cerroni, C. W. (2004). XPeer: A Self-organizing XML P2P Database System. *First EDBT Workshop on P2P and Databases (P2P&DB)*.
- Schickel, V., & Faltings, B. (2007). Oss: A semantic similarity function based on hierarchical ontologies. *IJCAI*, (σσ. 551-556).
- Schlosser, M., Sintek, M., Decker, S., & Nejdl, W. (2002). Hypercup hypercubes, ontologies and efficient search on p2p networks. *International Workshop on Agents and Peer-to-Peer Computing*. Bologna, Italy.
- Schlosser, M., Sintek, M., Decker, S., & Nejdl, W. (2002). Hypercup- hypercubes, ontologies, and efficient search on peer-to-peer networks. In: Moro G, Koubarakis M,. (AP2PC, Επιμ.) *Lecture notes in computer science*, 2530, 112–124.
- Schmitz, C., & Loser, A. (2006). How to model Semantic Peer-to-Peer Overlays? *P2PIR Workshop, Informatik*.
- Selinger, G., Astrahan, M., Chamberlin, D., Lorie, R., & Price, T. (1979). Access Path Selection in a Relational Database Management System. *ACM*.
- Shannon, C. E. (1948, October). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27, 379-423.
- Soldak, Y. (2005). *Spring Young Researcher's Colloquium on Database and Information Systems SYRCoDIS*. St.-Petersburg, Russia.
- Sourceforge.net. (n.d.). FADA. Ανάκτηση από <http://fada.sourceforge.net/>
- Stocker, M., Seaborne, A., & Bernstein, A. (2008). SPARQL basic graph pattern optimization using selectivity estimation. *WWW*. ACM.
- Stoica, I., Morris, R., Karger, D., Kaashoek, M., & Balakrishnan, H. (2001). Chord: a scalable peerto-peer lookup service for Internet applications. *ACM SIGCOMM 01*.
- Stoica, I., Morris, R., Liben-Nowell, D., Karger, D. R., Frans Kaashoek, M., Dabek, F., και συν. (2003). Chord: A Scalable Peer-to-peer Lookup Protocol for Internet Applications. *IEEE/ACM Transactions on Networking*, 11(1), 17-32.
- Stoica, I., Morris, R., Liben-Nowell, D., Karger, D. R., Frans Kaashoek, M., Dabek, F., και συν. (February 2003). Chord: A Scalable Peer-to-peer Lookup Protocol for Internet Applications. *IEEE/ACM Transactions on Networking*, 11(1), 17-32.
- Tang, C., & Dwarkadas, S. (2004). Hybrid Global-Local Indexing for Efficient Peer-to-Peer Information Retrieval. *NSDI*.
- Tang, C., Xu, Z., & Dwarkadas, S. (2003). Peer-to-peer information retrieval using self- organizing semantic overlay networks. *SIGCOMM*. Karlsruhe, Germany: ACM.

- Tao Shen, H., Shu, Y., & Yu, B. (2004, July). Efficient Semantic-Based Content Search in P2P Networks. *IEEE Transactions on Knowledge and Data Engineering*, 16(7), 813-826.
- Tempich, C., Staab, S., & Wranik, A. (2004). Remindin': semantic query routing in peer-to-peer networks based on social metaphors. *13th international conference on World Wide Web (WWW'04)* (σσ. 640–649). ACM Press.
- Thanh Le, B., Dieng-Kuntz, R., & Gandon, F. (2004). On ontology matching problems for building a corporate Semantic Web in a multi-communities organization. *ICEIS*.
- The Apache Derby Project*. (n.d.). Ανάκτηση από <http://incubator.apache.org/derby/>
- Verma, K. (n.d.). METEOR-S WSDI: A Scalable Infrastructure of Registries for Semantic Publication and Discovery of Web Services. *Journal of Information Technology and Management*.
- Voulgaris, S., Kermarrec, A.-M., Massoulie, L., & van Steen, M. (2004). Exploiting semantic proximity in peer-to-peer content searching. *10th international workshop on future trends in distributed computing systems (FTDCS)*. Suzhou, China: <http://www.cs.vu.nl/spyros/papers/ftdcs.04.pdf>.
- Wang, Y., Wang, H., Meng, X., & Wang, S. (2004). Estimating the Selectivity of XML Path Expression with Predicates by Histograms. *WAIM* (σσ. 409-418). Springer-Verlag.
- Yang, B., & Garcia-Molina, H. (2001). Comparing hybrid peer-to-peer systems. *VLDB*.
- Yang, B., & Garcia-Molina, H. (2003). Designing a Super-peer Network. *19th International Conference on Data Engineering, ICDE*.
- Zhao, B., Kubiawicz, J., & Joseph, A. (April 2001). *Tapestry: An infrastructure for fault-tolerant wider-area location and routing*. U.C. Berkeley: Technical Report UCB/CSD-0101141, Computer Science Division.
- Zhao, Y. B., Kubiawicz, J., & Joseph, A. D. (2001). *Tapestry: an infrastructure for fault-tolerant wide-area location and routing*. Technical report, University of California, Berkeley.

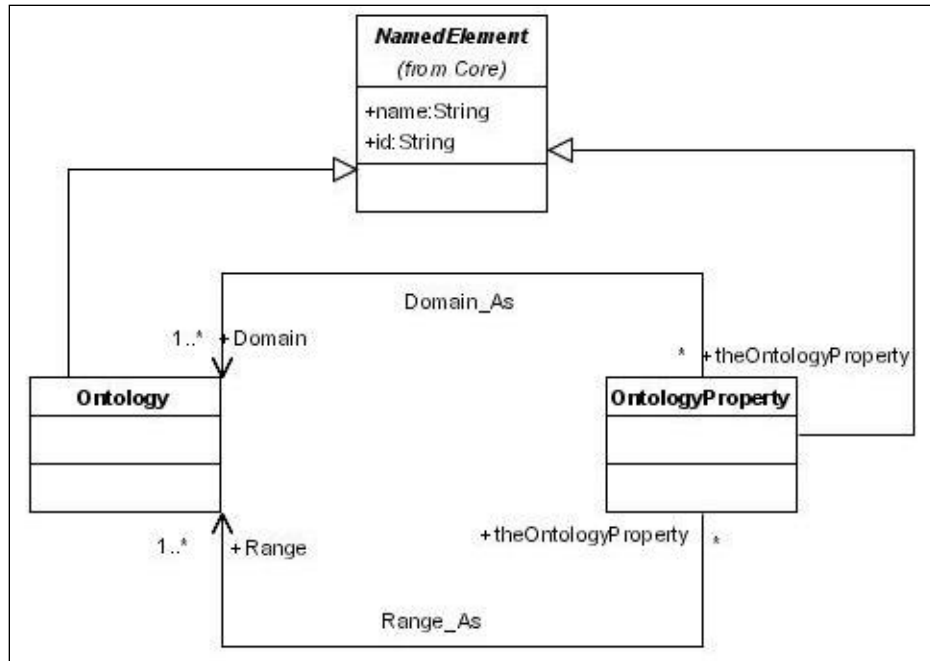
Παράρτημα Α - Η γλώσσα περιγραφής οντολογιών ODM



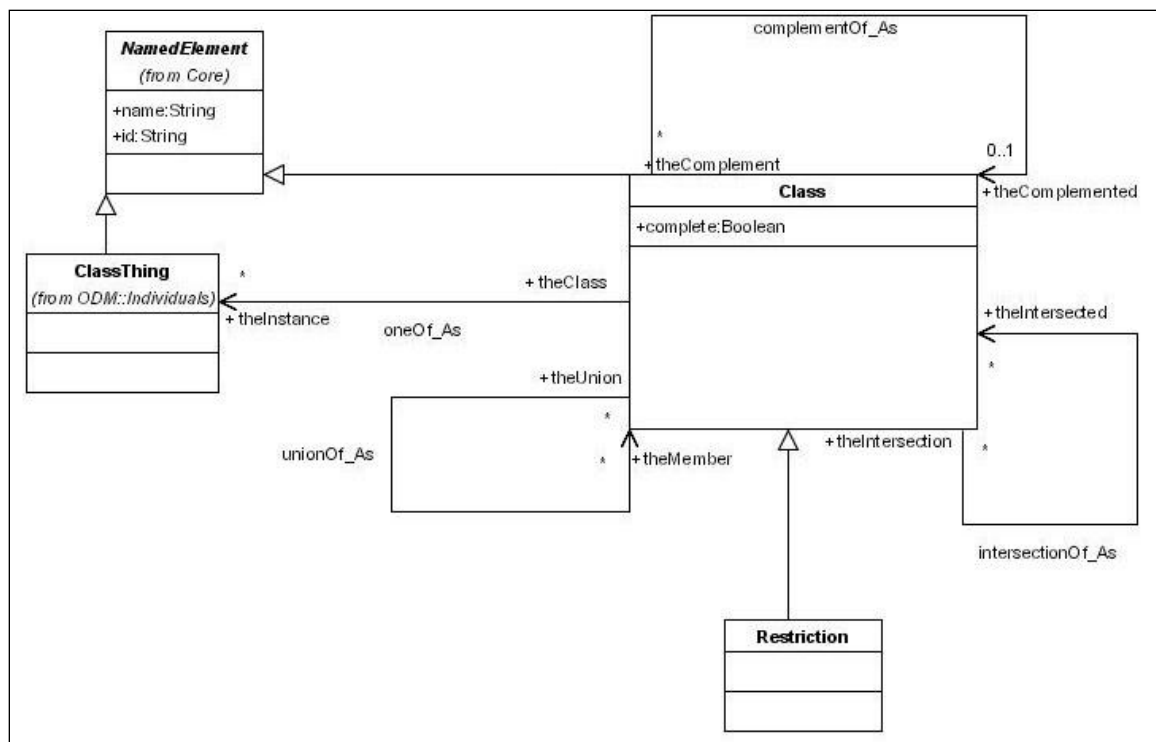
Σχήμα 24 Package View of the ODM



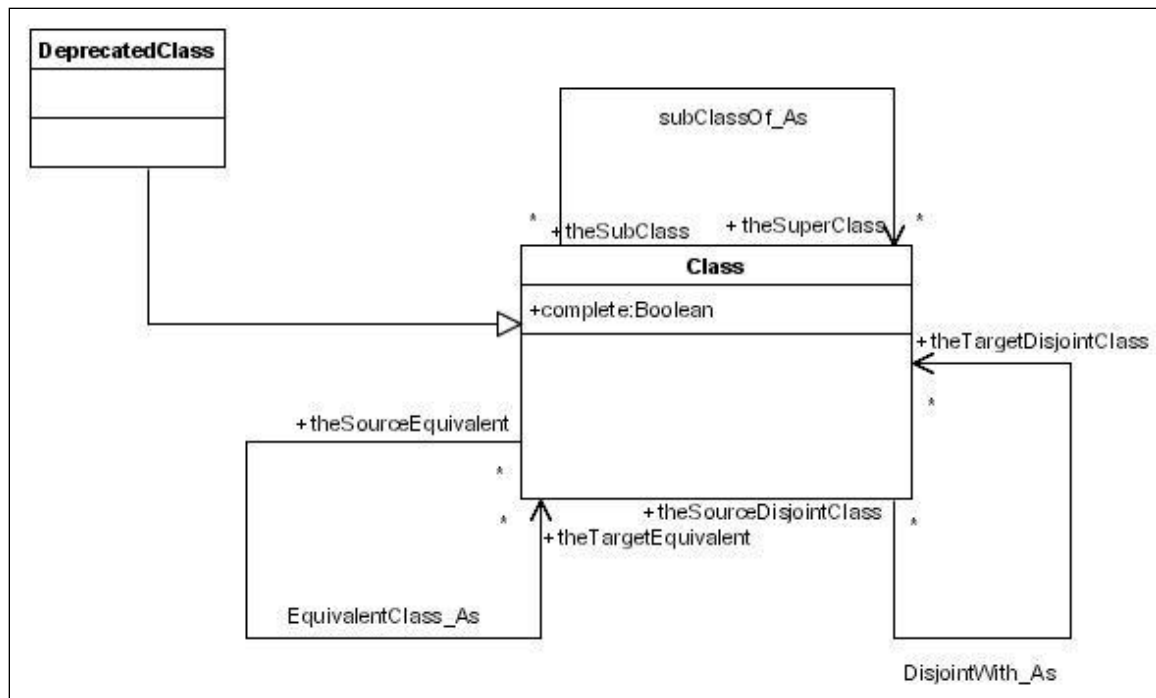
Σχήμα 25 ODM Abstract Concepts



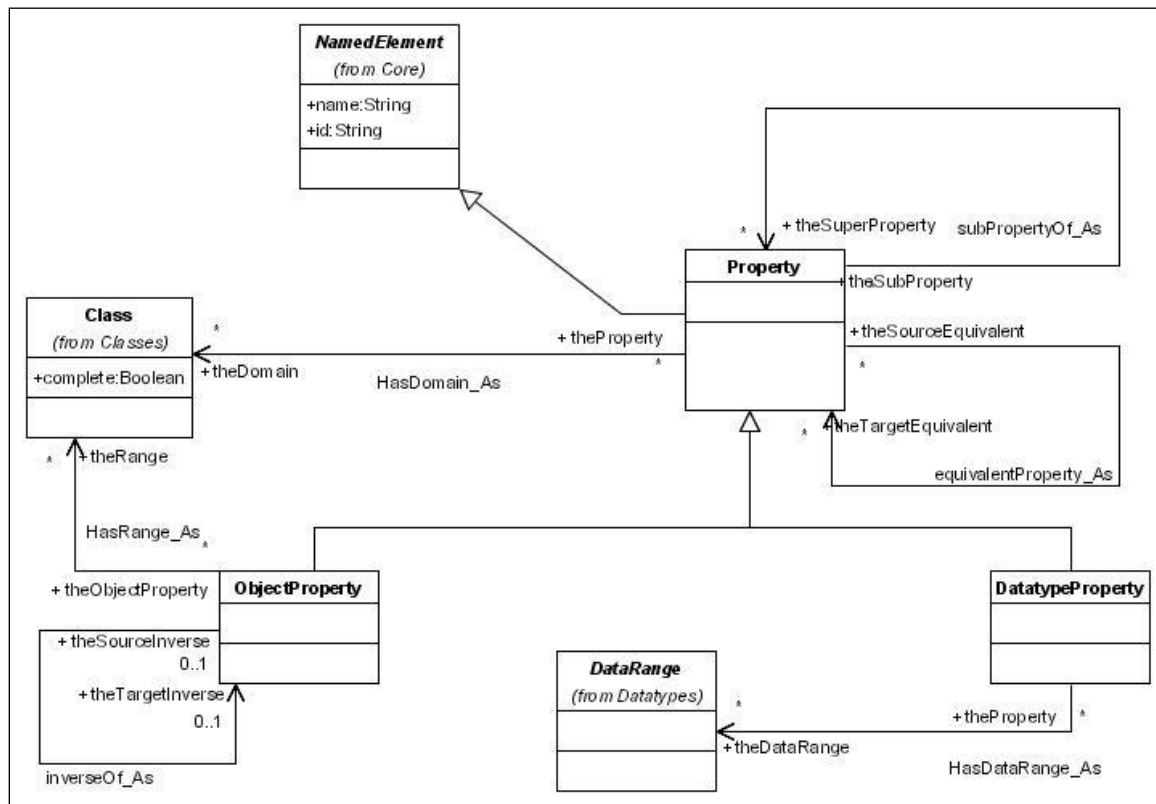
Σχήμα 26 Ontology and Ontology Properties



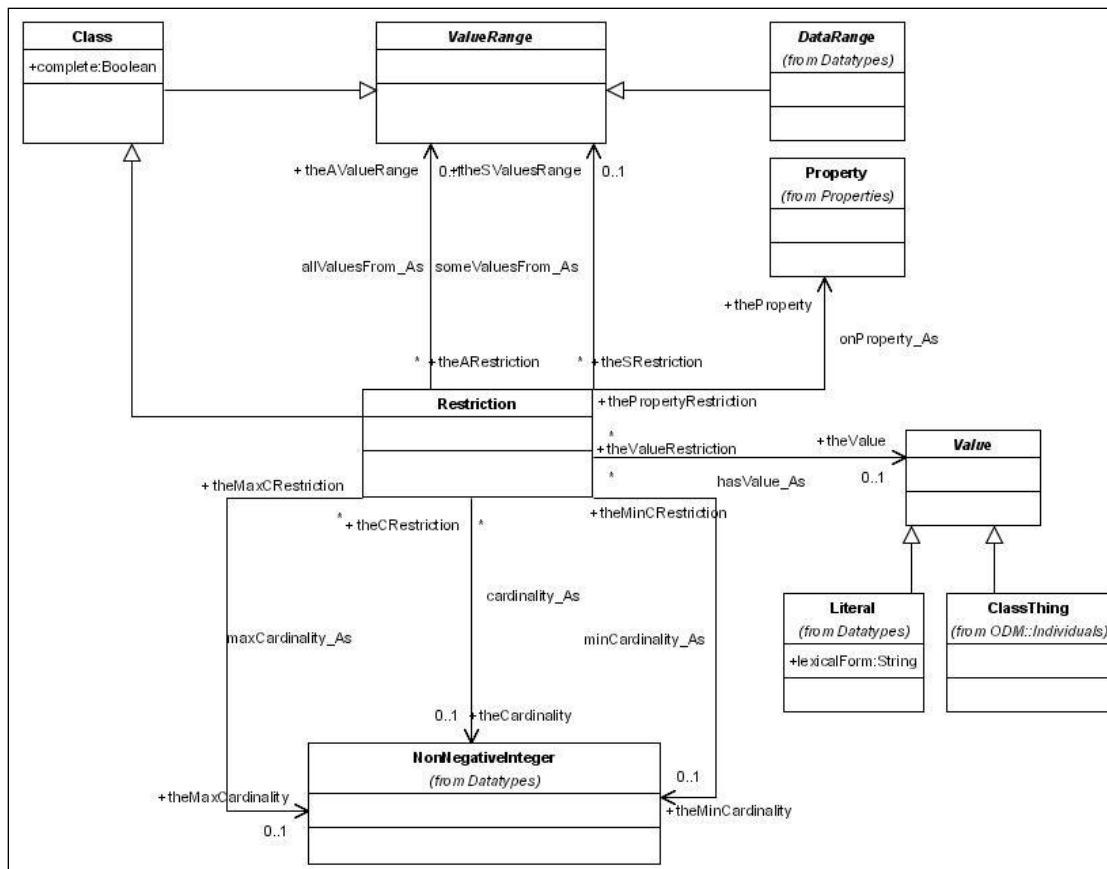
Σχήμα 27 ODM Class Definitions



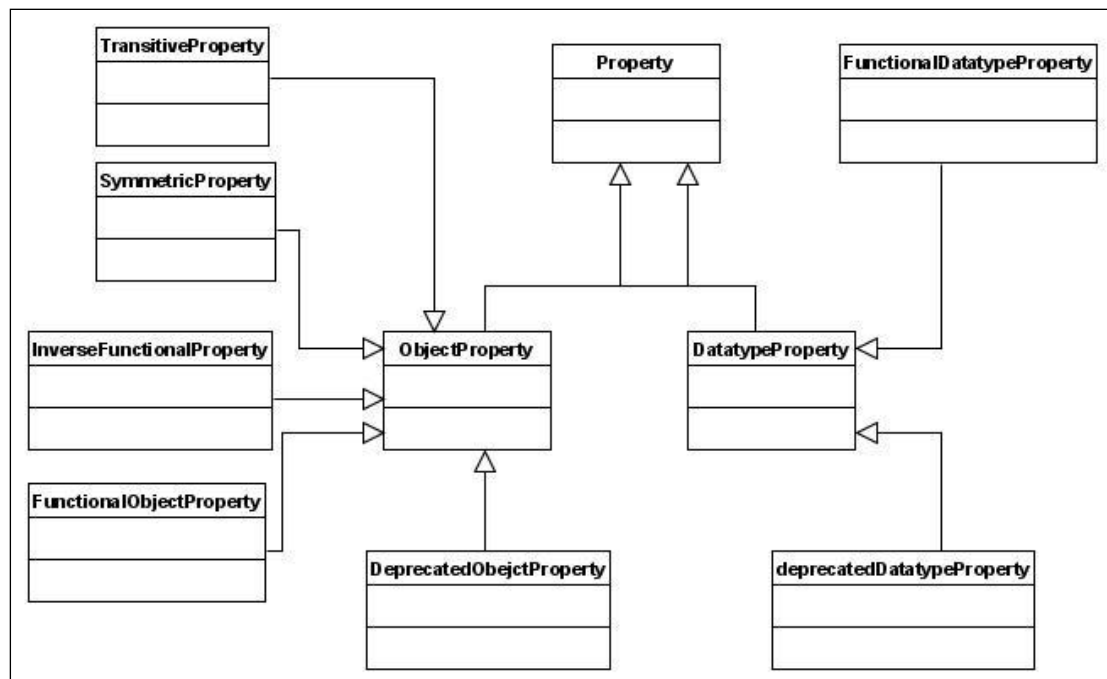
Σχήμα 28 ODM Class Axioms



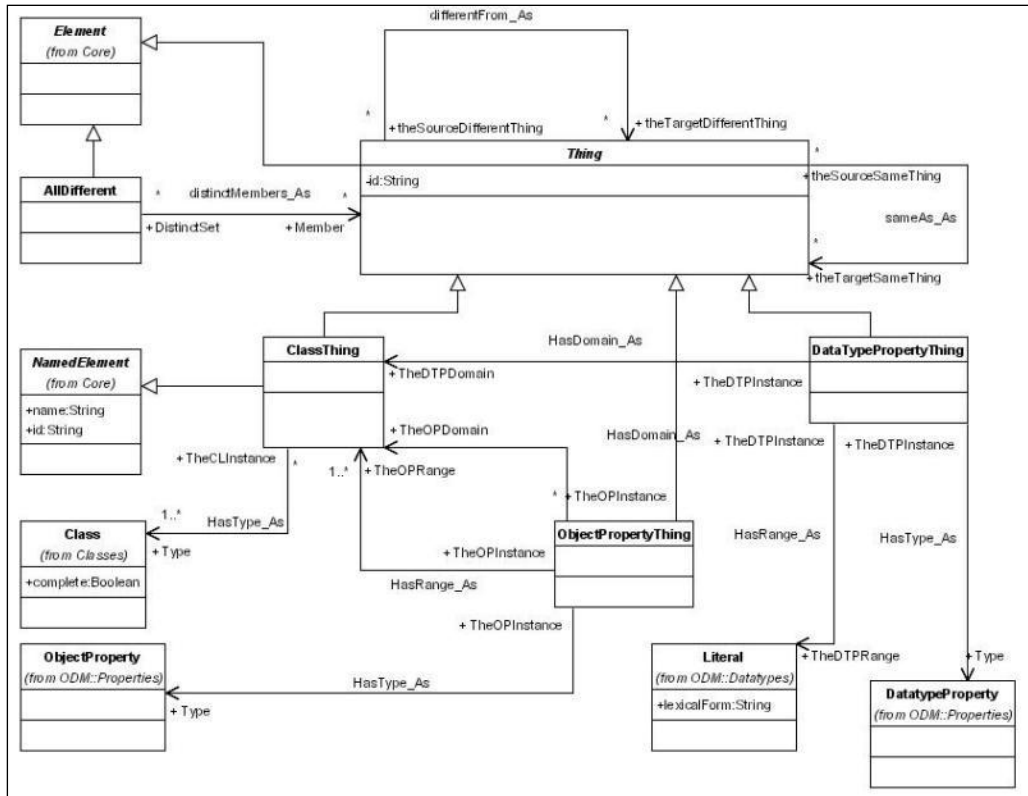
Σχήμα 29 Properties in ODM



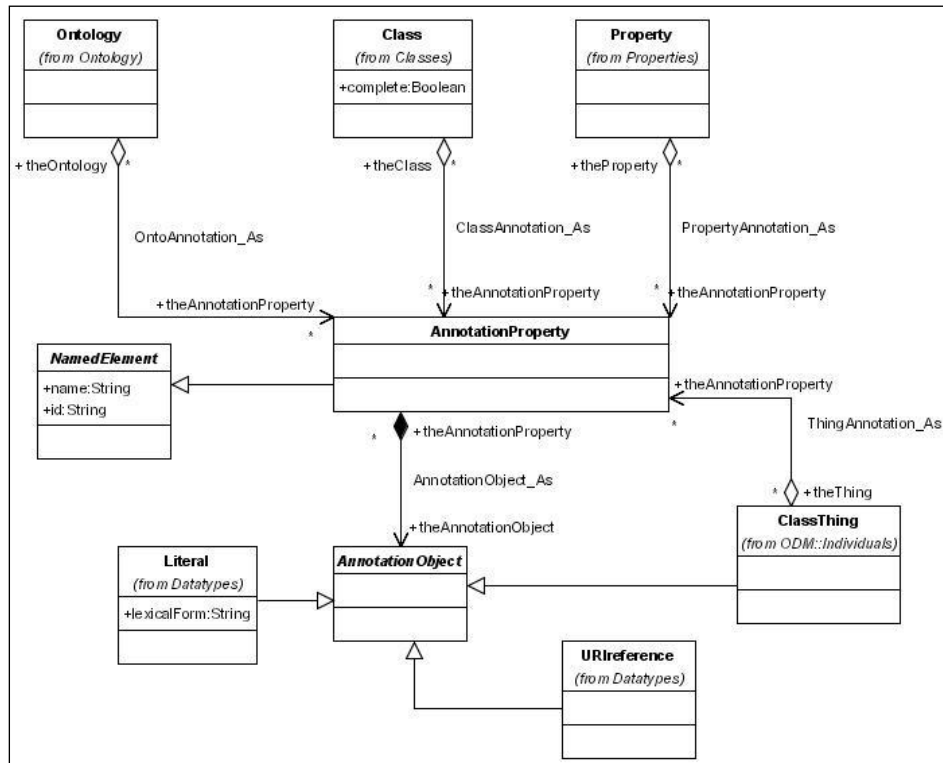
Σχήμα 30 ODM Property Restrictions



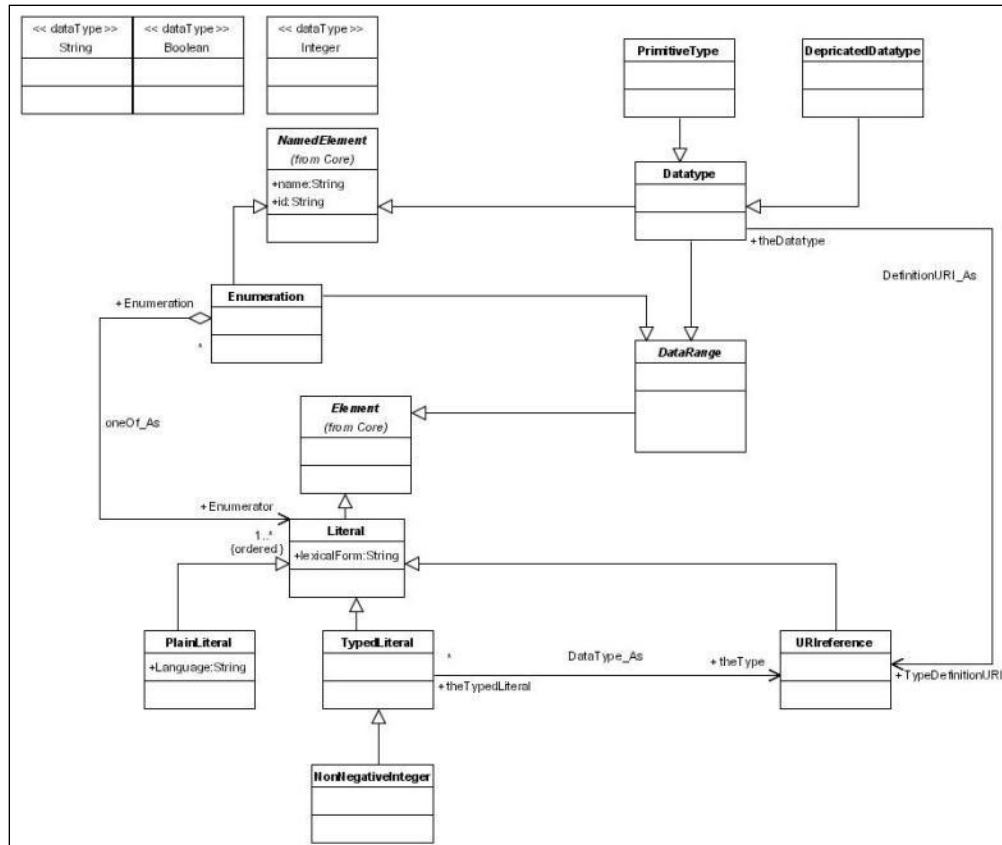
Σχήμα 31 ODM Constructs for Property Axioms



Σχήμα 32 Individuals in ODM



Σχήμα 33 ODM Annotations



Σχήμα 34 ODM Data Types