ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αυτόματη Συμπλήρωση Οντολογίας Προϊόντων από Σελίδες του Παγκόσμιου Ιστού

Μαρινέλα Β. Μάρκο

Επιβλέποντες: Ευστάθιος Χατζηευθυμιάδης, Επίκουρος Καθηγητής ΕΚΠΑ
Κωνσταντίνος Κολομβάτσος, Υποψήφιος Διδάκτωρ ΕΚΠΑ

ΑΘΗΝΑ

ΔΕΚΕΜΒΡΙΟΣ 2011

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αυτόματη Συμπλήρωση Οντολογίας Προϊόντων από Σελίδες του Παγκόσμιου Ιστού

Μαρινέλα Β. Μάρκο

Α.Μ.: M1048

ΕΠΙΒΛΕΠΩΝΤΕΣ:

Ευστάθιος Χατζηευθυμιάδης, Επίκουρος Καθηγητής ΕΚΠΑ
Κωνσταντίνος Κολομβάτσος, Υποψήφιος Διδάκτωρ ΕΚΠΑ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Δεκέμβριος 2011

ΠΕΡΙΛΗΨΗ

Στις σύγχρονες εφαρμογές του Σημασιολογικού Ιστού, η εξέλιξη των οντολογιών γίνεται όλο και πιο σημαντική ως εργαλείο υποστήριξης στην διαχείριση των οντολογιών, μέσω προηγμένων και ενδεχομένως αυτόματων τεχνικών. Μια από τις κύριες δραστηριότητες στην εξέλιξη των οντολογιών είναι η συμπλήρωση οντολογίας (ontology population), όπου η οντολογία αναπτύσσεται αποκτώντας νέες σημασιολογικές περιγραφές των δεδομένων και στιγμιότυπα που εξάγονται από ετερογενείς πηγές δεδομένων.

Στο πλαίσιο αυτό, η παρούσα εργασία προσπαθεί να επεκτείνει μια υπάρχουσα οντολογία προϊόντων με νέα στιγμιότυπα προϊόντων που εξάγονται από αντίστοιχες HTML σελίδες ηλεκτρονικών καταστημάτων. Στο σύστημα που αναπτύχθηκε τα προϊόντα εισάγονται στις κατάλληλες κατηγορίες που έχουν οριστεί στην οντολογία. Για τον σκοπό αυτό χρησιμοποιούνται τεχνικές σημασιολογικής και λεξικογραφικής ομοιότητας. Στόχος είναι να διαπιστώσουμε την ομοιότητα μεταξύ κατηγοριών προϊόντων στην ιστοσελίδα με αυτών στην οντολογία, ώστε να επιτευχθεί η συμπλήρωση με νέα προϊόντα.

Τα αποτελέσματα αξιολόγησης του συστήματος είναι θετικά όσο αφορά στην ανάκτηση των προϊόντων και των χαρακτηριστικών τους, και στην περαιτέρω αποθήκευσή τους στην οντολογία. Ωστόσο, η μέθοδος αποθήκευσης της κατηγορίας προϊόντων στην κατάλληλη θέση στην οντολογία χρήζει περαιτέρω βελτίωσης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Σημασιολογικός Ιστός

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Συμπλήρωση Οντολογίας Προϊόντων, Σημασιολογικός Ιστός, Οντολογίες, Προιόντα, Ηλεκτρονικές Αγορές, HTML, OWL

ABSTRACT

In modern Semantic Web applications, ontology evolution is becoming more important. This is due to the need of supporting experts in managing ontology changes through advanced, and possibly automated, techniques. One of the main activities in ontology evolution is ontology population, where the ontology is evolved by acquiring new product semantic descriptions of data extracted from heterogeneous data sources.

This thesis attempts to automatically populate an existing product ontology with new instances that are extracted from corresponding online shops' HTML pages. More specifically, the methodology adopted focuses on HTML elements handling (i.e. tables or lists). In the system developed, the products are imported into the appropriate categories which are defined in the ontology. To achieve this, there have been adopted semantic and lexicographical similarity techniques. The aim is to find similarity between product categories in the website with those existing in the ontology, so as to implement the ontology population process.

The evaluation results of the system are positive, regarding the retrieval of products and their characteristics. However, the methodology used to store the product category in the appropriate position in the ontology needs further enhancement.

SUBJECT AREA: Semantic Web

KEYWORDS: Semantic Web, Ontologies, Ontology Population, Products, Electronic Marketplaces, HTML, OWL

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της εργασίας, επίκουρο καθηγητή κ.Ευστάθιο Χατζηευθυμιάδη για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα θέμα που πραγματικά με ενδιέφερε. Επίσης, ευχαριστώ ειλικρινώς τον υποψήφιο διδάκτορα κ. Κωνσταντίνο Κολομβάτσο, χωρίς την διαρκή βοήθεια και ανατροφοδότηση του οποίου δεν θα είχε ολοκληρωθεί αυτή η εργασία.

Τέλος, ευχαριστώ την οικογένειά μου για όλα όσα μου έχει προσφέρει ως σήμερα.

ΠΕΡΙΕΧΟΜΕΝΑ

ΛΙΣΤΑ ΣΧΗΜΑΤΩΝ – ΕΙΚΟΝΩΝ

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Γενικά

Η αποτελεσματική επικοινωνία μεταξύ μηχανών είναι μια απαίτηση των εικονικών αγορών και ο Σημασιολογικός Ιστός υπόσχεται να κάνει τον μεγάλο όγκο των δεδομένων που υπάρχουν στον Ιστό, αναγνώσιμο από τις μηχανές και επεξεργάσιμο μέσω της τυποποίησης της σημασιολογίας του. Εφόσον, τα προϊόντα και οι υπηρεσίες είναι τα βασικά αντικείμενα εμπορίου, η αναπαράστασή τους σε μορφή που θα καταλαβαίνουν οι μηχανές είναι μια πρόκληση κλειδί στην πορεία για επιχειρηματικές εφαρμογές για τον Σημασιολογικό Ιστό. Οι οντολογίες αποτελούν την βασική τεχνολογία για την πραγματοποίηση αυτής της προσπάθειας.

Μια οντολογία χρησιμοποιείται ευρέως σε πολλών ειδών εφαρμογές ως ένα εργαλείο για αναπαράσταση γνώσης του πεδίου ενδιαφέροντος. Όμως, αν και η οντολογία δημιουργείται προσεκτικά από τους εμπειρογνώμονες, η προσθήκη στιγμιοτύπων (instances) είναι κουραστική και δαπανηρή. Η χειροκίνητη διαδικασία για προσθήκη στιγμιοτύπων προβλέπει ότι κάποιος πρέπει να βρεί τις πληροφοριακές πηγές που περιλαμβάνουν δεδομένα για τις έννοιες (concepts) της οντολογίας, να ταξινομήσει την έννοια σύμφωνα με τα περιεχόμενά της και να συγχωνεύσει πολλών ειδών πληροφορίες ως στέρεη-τελική γνώση. Ως αντίδοτο σε αυτή την πολύπλοκη διαδικασία, πολλοί ερευνητές τελευταία έχουν στρέψει την προσοχή τους στην αυτοματοποίησή της. Ο όρος *αυτόματη συμπλήρωση οντολογίας (automatic ontology population)* χρησιμοποιείται γενικά για αυτή την διαδικασία.

Η αυτόματη εισαγωγή στιγμιοτύπων σε μια οντολογία από αδόμητο κανονικό κείμενο είναι πρακτικά αδύνατη λόγω της κακής επίδοσης των διαφόρων τεχνικών επεξεργασίας φυσικής γλώσσας [23]. Ο προσδιορισμός στιγμιοτύπων από προτάσεις φυσικής γλώσσας απαιτεί συντακτική ανάλυση, αποσαφήνιση της

έννοιας των λέξεων, αναγνώριση των ονομάτων-οντοτήτων, εξαγωγή συσχετίσεων, coreference solution κτλ. Όμως, όλες αυτές οι τεχνικές υποφέρουν από σοβαρές ασάφειες των οποίων η επίλυση δεν είναι ακόμη ικανοποιητική. Όταν η ανάλυση της φυσικής γλώσσας είναι ανεπαρκής, η υποψήφια πηγή πληροφορίας είναι οι σελίδες Ιστού, τα ημιδομημένα έγγραφα.

Συγκεκριμένα, οι σελίδες Ιστού περιλαμβάνουν έναν αριθμό από πίνακες και HTML λίστες, τα οποία έχουν μια οργανωμένη δομή παρουσίασης. Εφόσον είναι σχετικά εύκολο να εξαχθεί γνώση από πίνακες και HTML λίστες, είναι πρακτικό να προστεθούν στιγμιότυπα σε οντολογία από σελίδες Ιστού που παρουσιάζουν τα προϊόντα τους σε πίνακες και λίστες.

Λαμβάνοντας υπόψη τα παραπάνω, το αντικείμενο μελέτης της παρούσας εργασίας στράφηκε στην εκμετάλλευση της συσσωρευμένης πληροφορίας για προϊόντα, που είναι ευρέως διαθέσιμη μέσω των σελίδων των ηλεκτρονικών καταστημάτων στον Ιστό.

1.2 Αντικείμενο και Στόχοι της Εργασίας

Το βασικό αντικείμενο μελέτης της εργασίας είναι η ανάπτυξη μιας μεθοδολογίας για αυτόματη εξαγωγή προϊόντων και χαρακτηριστικών τους από HTML λίστες και πίνακες σελίδων ηλεκτρονικών καταστημάτων, και η περαιτέρω αποθήκευσή τους σε μια υπάρχουσα οντολογία προϊόντων. Οι σελίδες των ηλεκτρονικών καταστημάτων είναι πλούσιες σε πληροφορίες για προϊόντα, τις οποίες παρουσιάζουν σε δομημένη μορφή μέσω HTML πινάκων ή λιστών. Αυτή την δομημένη μορφή παρουσίασης των προϊόντων προσπαθεί η εκμεταλλευτεί η παρούσα εργασία, όπου κάνοντας λεξιλογική ανάλυση των τιμών των HTML χαρακτηριστικών (attributes) των στοιχείων πινάκων και λιστών των ιστοσελίδων, προσπαθεί να εξάγει τα χαρακτηριστικά των προϊόντων και να τα εισάγει ως ιδιότητες στα στιγμιότυπα προϊόντων στην οντολογία. Η διαδικασία αυτή απαιτεί αρχικά την εύρεση της σωστής κλάσης στην οντολογία που είναι η κατηγορία στην οποία θα εισαχθούν τα προϊόντα, όπου σε περίπτωση μη ύπαρξης της

κατηγορίας στην οντολογία δημιουργείται νέα κλάση-κατηγορία στην ιεραρχία κλάσεων της οντολογίας.

Ο στόχος για την υλοποίηση αυτής της προσέγγισης είναι η δημιουργία μιας εμπλουτισμένης οντολογίας προϊόντων που θα περιλαμβάνει χαρακτηριστικά προϊόντων όπως είναι οι τιμές, οι πάροχοι, η κατηγορία, ο κατασκευαστής, η χώρα προέλευσης κτλ.. Η εμπλουτισμένη οντολογία θα μπορεί να χρησιμοποιηθεί για επεξεργασία από άλλες οντότητες (ανθρώπινες ή τεχνητές) για την επίτευξη πιο πολύπλοκων στόχων. Για παράδειγμα, ένα πιθανό σενάριο χρήσης της εμπλουτισμένης οντολογίας προϊόντων είναι στο περιβάλλον μιας εικονικής αγοράς. Η εμπλουτισμένη οντολογία προϊόντων θα μπορούσε να χρησιμοποιηθεί από μια αυτόματη μηχανή αγοράς (shopbot) για να κάνει επερωτήσεις με βάση συγκεκριμένα χαρακτηριστικά προϊόντων, ώστε να δωθεί στον χρήστη προσωποιημένη πληροφόρηση, χωρίς να χρειάζεται η πλοήγηση και σάρωση σε όλες τις HTML σελίδες των ηλεκτρονικών καταστημάτων από την αρχή, καθώς η πληροφορία για τα προϊόντα θα είναι αποθηκευμένη στην οντολογία προϊόντων.

1.3 Οργάνωση Κεφαλαίων

Τα περιεχόμενα της παρούσας εργασίας κατανέμονται σε 7 κεφαλαία.

Το Κεφάλαιο 2 ορίζει την έννοια του Σημασιολογικού Ιστού και παρουσιάζει τις κυριότερες τεχνολογίες που το απαρτίζουν. Επίσης, παρουσιάζονται θέματα που αφορούν στην αναπαράσταση προϊόντων στο περιβάλλον του Σημασιολογικού Ιστού. Στο δεύτερο μέρος του κεφαλαίου ορίζονται οι οντολογίες του Σημασιολογικού Ιστού και η γλώσσα διαχείρισης οντολογιών OWL. Επίσης, περιγράφεται η έννοια της μάθησης οντολογιών και οι διαφορετικές προσεγγίσεις μάθησης οντολογίας που υπάρχουν σήμερα.

Στο Κεφάλαιο 3 παρουσιάζονται οι εικονικές αγορές, τα χαρακτηριστικά, τα οφέλη που παρέχουν, οι οντότητες που το απαρτίζουν καθώς και οι βασικές κατηγοριοποιήσεις τους. Επίσης, γίνεται ανάλυση του ρόλου των οντολογιών στις

εικονικές αγορές και ποια η συνεισφορά τους σε αυτές. Τέλος, παρουσιάζονται οι τεχνικές που υπάρχουν σήμερα για μάθηση οντολογιών από σχεσιακά σχήματα, τα οποία αποτελούν την πηγή αποθήκευσης των δεδομένων προϊόντων που είναι διαθέσιμα στις εικονικές αγορές.

Στο Κεφάλαιο 4 γίνεται παρουσίαση των διαθέσιμων οντολογιών προϊόντων και των προτύπων κατηγοριοποίησης eCl@ss και UNSPSC. Στη συνέχεια αναφέρονται συμπεράσματα σχετικά με την χρήση των προτύπων αυτών αλλά και γενικά για την κατηγοριοποίηση των προϊόντων. Επίσης τονίζεται η ανάγκη για χρήση μιας πιο light οντολογίας ιδιαίτερα δε όταν πρόκειται να χρησιμοποιηθεί από ένα shopbot.

Στο Κεφάλαιο 5 γίνεται εκτενής παρουσίαση της αρχιτεκτονικής του συστήματος που αναπτύχθηκε, των σχεδιαστικών επιλογών και μεθόδων υλοποίησης, καθώς και της οντολογίας προϊόντων που χρησιμοποιήθηκε. Επίσης, αναφέρονται οι τεχνολογίες υλοποίησης του συστήματος.

Στο Κεφάλαιο 6 γίνεται αξιολόγηση του συστήματος και παρουσιάζονται τα αποτελέσματα των σεναρίων εκτέλεσης. Επίσης, γίνεται ανάλυση των ποιοτικών και ποσοτικών επιδόσεων της εφαρμογής αναφορικά με τα σενάρια εκτέλεσης.

Τέλος, στο Κεφάλαιο 7 παρουσιάζονται τελικά συμπεράσματα και πιθανές μελλοντικές κατευθύνσεις βελτιστοποίησης του συστήματος.

ΚΕΦΑΛΑΙΟ 2

ΣΗΜΑΣΙΟΛΟΓΙΚΟΣ ΙΣΤΟΣ ΚΑΙ ΟΝΤΟΛΟΓΙΕΣ

2.1 Σημασιολογικός Ιστός

Ο Σημασιολογικός Ιστός (Semantic Web) αποτελεί μια εξέλιξη του Παγκόσμιου Ιστού (World Wide Web), όπου παρέχοντας μια καλά ορισμένη σημασιολογία στις πληροφορίες, επιτρέπεται στις μηχανές να «καταλαβαίνουν» τη σημασία της υπερσυνδεδεμένης πληροφορίας. Αυτό καθιστά δυνατό οι άνθρωποι να μοιράζονται περιεχόμενα πέρα από τα όρια συγκεκριμένων εφαρμογών και ιστοσελίδων. Έχει περιγραφεί με αρκετά διαφορετικούς τρόπους, ως μια ουτοπική προοπτική, ως ένα δίκτυο δεδομένων ή απλά ως μια φυσική μεταστροφή στον καθημερινό τρόπο χρήσης του Ιστού.

Το όραμα του Σημασιολογικού Ιστού όπως έχει περιγραφεί από τον εμπνευστή του, Tim Berners-Lee [1], λέει ότι:

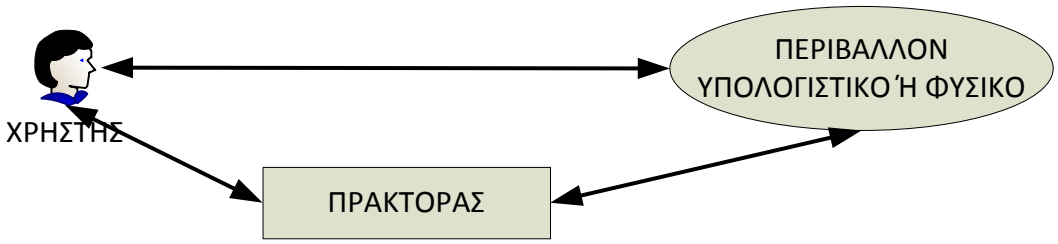
"*Ο Σημασιολογικός Ιστός θα δώσει δομή στη σημασιολογία των περιεχομένων των Ιστοσελίδων, δημιουργώντας ένα περιβάλλον όπου οι πράκτορες λογισμικού περιπλανόμενοι από σελίδα σε σελίδα θα μπορούν να εκτελούν πολύπλοκες εργασίες για τους χρήστες.*"

Οι μέθοδοι που πρέπει να χρησιμοποιηθούν για να υλοποιήσουμε τον Σημασιολογικό Ιστό είναι αντί να δημοσιεύονται πληροφορίες που να είναι κατανοητές μόνο από τον άνθρωπο, να γίνεται δημοσίευση δεδομένων και μετα-δεδομένων που μπορούν να επεξεργαστούν οι μηχανές, με τη χρήση όρων και γλωσσών που αυτές καταλαβαίνουν. Στη συνέχεια πρέπει να δημιουργηθούν πράκτορες που θα αναζητήσουν, θα κάνουν επερωτήσεις, θα ενοποιήσουν κτλ. αυτά τα δεδομένα. Τέλος, θα πρέπει να είναι βέβαιο ότι όλοι οι πράκτορες θα καταλαβαίνουν αυτούς τους όρους / γλώσσες.

Υπάρχει αρκετή βιβλιογραφία για τους πράκτορες. Περιληπτικά αναφέρουμε ότι ένας *πράκτορας (agent)* είναι μια οντότητα που αντιλαμβάνεται το *περιβάλλον* μέσα στο οποίο βρίσκεται με τη βοήθεια *αισθητήρων (sensors)*, είναι μέρος του

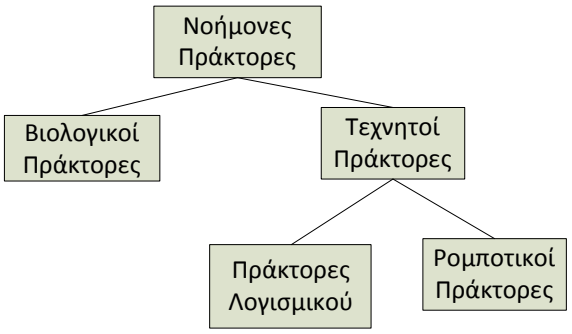
περιβάλλοντος αυτού, κάνει συλλογισμούς για το περιβάλλον και δρα πάνω σε αυτό με τη βοήθεια *μηχανισμών δράσης (effectors)*, για την επίτευξη κάποιων *στόχων* [24].

Ο παραπάνω ορισμός εμπεριέχει την έννοια της αυτονομίας ενός πράκτορα, δηλαδή της αυτενέργειας για την επίτευξη των στόχων του. Αυτό το χαρακτηριστικό είναι ίσως ο πιο κοινός παρονομαστής όλων των ειδών πρακτόρων, και υποχρεώνει την ύπαρξη “νοημοσύνης”, τουλάχιστον σε κάποιο βαθμό (intelligent agents).



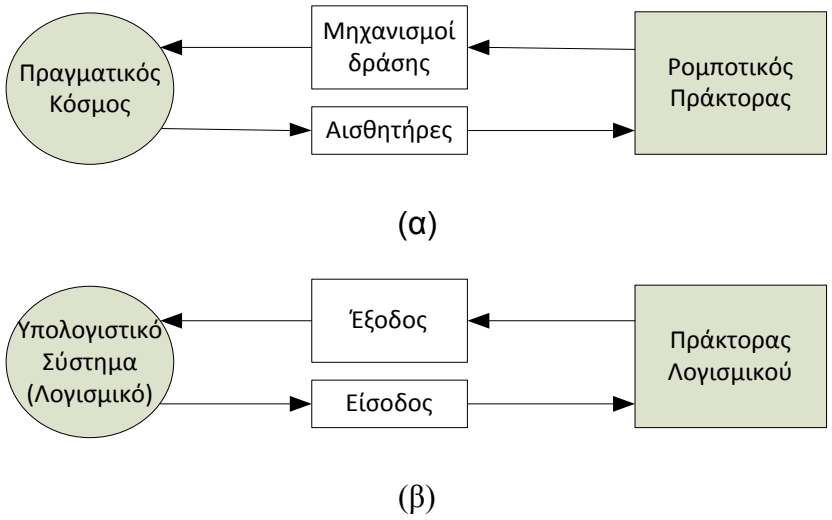
Σχήμα 2.1 : Διασύνδεση χρήστη-προγράμματος μέσω πρακτόρων.

Η τεχνολογία των πρακτόρων αλλάζει τη μορφή της διασύνδεσης χρήστη-λογισμικού. Ο χρήστης έχει την δυνατότητα να μην επικοινωνεί απευθείας με κάποια εφαρμογή αλλά να χρησιμοποιεί έναν πράκτορα ο οποίος τον διευκολύνει σε χρονοβόρες διαδικασίες, διαδικασίες ρουτίνας ή διαδικασίες που χρειάζονται κάποια ικανότητα που ο χρήστης δεν έχει αποκτήσει ακόμη (Σχήμα 2.1).



Σχήμα 2.2 : Μια ταξινόμηση των πρακτόρων.

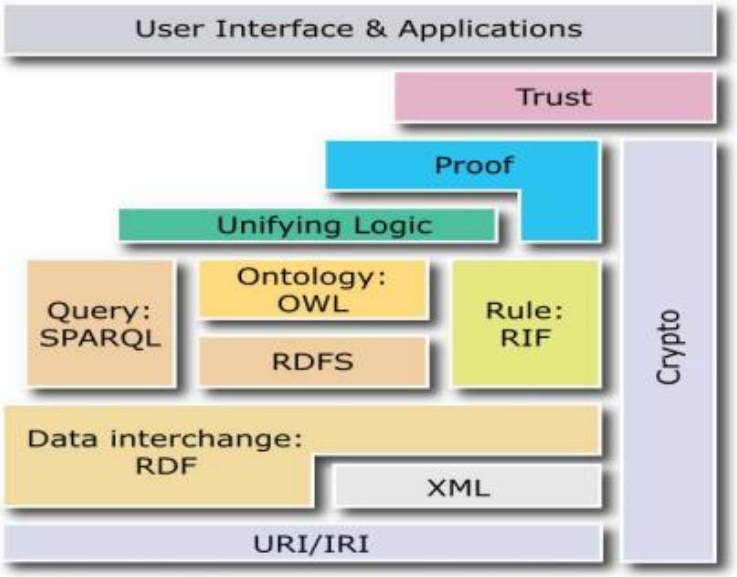
Ένας αφηρημένος διαχωρισμός των πρακτόρων είναι μεταξύ βιολογικών και τεχνητών (Σχήμα 2.2). Παραπέρα, οι τεχνητοί πράκτορες διαχωρίζονται σε *ρομποτικούς (robotic agents ή robots)*(Σχήμα 2.3(α)), οι οποίοι έχουν σαν αισθητήρες και μηχανισμούς δράσης μηχανικά ή ηλεκτρονικά μέρη και δρουν στον πραγματικό κόσμο, και *πράκτορες λογισμικού (software agents ή softbots)* (Σχήμα 2.3(β)), οι οποίοι είναι προγράμματα και δρουν σε ένα υπολογιστικό σύστημα.



Σχήμα 2.3 (α) Ρομποτικός πράκτορας, (β) Πράκτορας λογισμικού

Το όραμα του Σημασιολογικού Ιστού σήμερα [2] αφορά δυο πράγματα. Πρώτο, αφορά κοινά πρότυπα για την ενοποίηση και τον συνδυασμό δεδομένων που εξάγονται από διαφορετικές πηγές, όταν στον παραδοσιακό Ιστό η κύρια προσέγγιση αφορά στην ανταλλαγή εγγράφων. Δεύτερο, αφορά στη γλώσσα για την καταγραφή του πώς τα δεδομένα σχετίζονται με αντικείμενα του πραγματικού κόσμου. Αυτό επιτρέπει σε έναν άνθρωπο ή μια μηχανή να ξεκινήσει από μια βάση δεδομένων και να μεταφέρεται σε ένα ατελείωτο σύνολο βάσεων δεδομένων που δεν συνδέονται με καλώδια, αλλά επειδή αναφέρονται στο ίδιο πράγμα. Έτσι, τα δυο πράγματα που τονίζονται είναι η αυξανόμενη ανάγκη για ενοποίηση των δεδομένων Ιστού, όπου πολλές εφαρμογές, όπως το e-science, e-government κ.ά., το απαιτούν, αλλά και η ανάγκη για πρότυπες γλώσσες Ιστού που να εκφράζουν την κατανεμημένη σημασιολογία, ώστε οι πράκτορες λογισμικού να μην αναπτύσσονται μόνο για συγκεκριμένες εργασίες.

Η ανάπτυξη του Σημασιολογικού Ιστού προχωράει σε βήματα, όπου κάθε βήμα χτίζει ένα επίπεδο πάνω στο προηγούμενο (Σχήμα 2.4).



Σχήμα 2.4: Η αρχιτεκτονική του Σημασιολογικού Ιστού.

Οι τεχνολογίες που χρησιμοποιεί ο Σημασιολογικός Ιστός και απεικονίζονται στο Σχήμα 2.4 είναι οι εξής :

- τα **URI**s (Universal Resource Identifier) : συμβολοσειρές που ταυτοποιούν μοναδικά μία οντότητα (ένα Web site, μία ιδιότητα, έναν άνθρωπο, ένα πράγμα κλπ) [10].
- επεκτάσιμη γλώσσα σήμανσης **XML** : επιτρέπει στους χρήστες να προσθέτουν αυθαίρετη δομή στα έγγραφά τους, χωρίς να καθορίζει την σημασιολογία αυτής της δομής [11].
- τεχνολογία **RDF** : χρησιμοποιείται για την αναπαράσταση δεδομένων και ανταλλαγής γνώσης στο διαδίκτυο [3].
- **RDF Schema** : παρέχει αρχές μοντελοποίησης για την οργάνωση σε ιεραρχίες των αντικειμένων στον Ιστό. Βασικά στοιχεία είναι οι κλάσεις (classes) και οι ιδιότητες (properties), οι σχέσεις υποκλάσης (subclass) και υπο-ιδιότητας (subproperty) και οι περιορισμοί πεδίου ορισμού (domain) και συνόλου τιμών (range). Το RDFS βασίζεται στο RDF. Μπορεί να θεωρηθεί ως μια πρωταρχική γλώσσα για τη σύνταξη οντολογιών. Όμως, χρειάζονται πιο ισχυρές γλώσσες οντολογιών που επεκτείνουν το RDFS και επιτρέπουν αναπαραστάσεις πιο πολύπλοκων σχέσεων μεταξύ αντικειμένων Ιστού [32].
- τεχνολογία **OWL** : χρησιμοποιείται για τη δημιουργία και διανομή οντολογιών, υποστηρίζοντας προχωρημένη αναζήτηση στο διαδίκτυο, πράκτορες λογισμικού και διαχείριση γνώσης [4].
- Γλώσσα επερωτήσεων **SPARQL** : Η γλώσσα SPARQL μπορεί να χρησιμοποιηθεί για να εκφράσει επερωτήσεις πάνω σε διαφορετικές πηγές δεδομένων, όταν αυτά είναι αποθηκευμένα σε RDF. Βασίζεται σε ταίριασμα graph pattern πάνω σε RDF γράφους. Ένα graph pattern είναι ένα σύνολο από τριπλέτες της μορφής (?x foaf:name ?name . ?x foaf:mbox ?mbox .). Τα αποτελέσματα των SPARQL επερωτήσεων μπορεί να είναι σύνολα ή RDF γράφοι [33].

- Η τεχνολογία **RIF** : Δημιουργήθηκε ως ένα πρότυπο για ανταλλαγή κανόνων μεταξύ συστημάτων κανόνων (rule systems), ειδικά μεταξύ μηχανών κανόνων στον Ιστό [34].

Αυτή η διαστρωματική ανάπτυξη του Σημασιολογικού Ιστού, πρέπει να ακολουθεί κάποιους κανόνες, ώστε να επιτυγχάνεται :

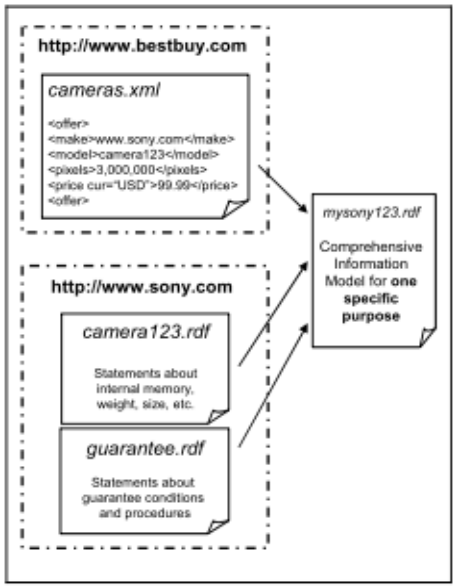
1. *Συμβατότητα προς τα κάτω* : Οι πράκτορες που έχουν πλήρη επίγνωση ενός επιπέδου, θα πρέπει επίσης να μπορούν να ερμηνεύουν και να χρησιμοποιούν πληροφορία που έχει γραφτεί σε χαμηλότερα επίπεδα. Για παράδειγμα, οι πράκτορες που γνωρίζουν την σημασιολογία της OWL μπορούν να επωφεληθούν πλήρως από την πληροφορία που έχει γραφτεί σε RDF και RDF Schema.
2. *Μερική κατανόηση προς πάνω* : Οι πράκτορες που έχουν πλήρη γνώση ενός επιπέδου, θα πρέπει να εκμεταλλευτούν μερικώς έστω την πληροφορία που υπάρχει σε υψηλότερα επίπεδα. Για παράδειγμα, ένας πράκτορας που γνωρίζει μόνο την σημασιολογία των RDF, RDFS μπορεί να ερμηνεύσει σε κάποιο βαθμό γνώση που είναι γραμμένη σε OWL, εάν δεν λάβει υπόψη τα στοιχεία που υπερβαίνουν τα RDF, RDFS.

2.1.1 Αναπαράσταση Προϊόντων στον Σημασιολογικό Ιστό

Στο περιβάλλον του Σημασιολογικού Ιστού, ο συμπερασμός πρόσθετης γνώσης από πολλαπλές εξωτερικές πηγές για την αντιμετώπιση της ελλιπούς πληροφορίας και της ανακρίβειας θα είναι κοινή πρακτική. Το ίδιο κοινές θα είναι και οι αλυσίδες επεξεργασίας πληροφοριών με πράκτορες λογισμικού να μεταφέρουν στοιχεία μεταξύ τους, ώστε να παράγουν το τελικό προϊόν ή υπηρεσία. Αυτή η αλυσιδωτή σύνδεση πληροφορίας από πολλές πηγές σημαίνει επίσης ότι η εγγενή δυναμική των εννοιών (π.χ. νέες κατηγορίες προϊόντων, νέα χαρακτηριστικά κ.ά.) από διαφορετικές σημασιολογικές κοινότητες ενδέχεται να πολλαπλασιαστεί. Κατά συνέπεια, η αναπαράσταση προϊόντων στον Σημασιολογικό Ιστό θα αντιμετωπίσει τρεις μεγάλες προκλήσεις [15]:

1. *Άγνωστος παραλήπτης δεδομένων* : Τα δεδομένα πρέπει να είναι κατάλληλα για ένα ευρύ κοινό και για ποικίλους σκοπούς, που μπορεί να είναι άγνωστες στον αρχικό εκδότη (Σχήμα 2.5). Στον Σημασιολογικό Ιστό, η πρόσβαση σε δημόσια έγγραφα θα είναι πολύ παρόμοια με τα τρέχοντα search engine bots που επισκέπτονται ιστοσελίδες μιας εταιρείας. Θα είναι πέρα από τον έλεγχο του παρόχου των δεδομένων ποιοι θα έχουν πρόσβαση και θα ερμηνεύουν αυτά τα δημόσια δεδομένα ή για ποιο σκοπό. Αξίζει να σημειωθεί ότι σήμερα, για Business to Business (B2B) ανταλλαγή δεδομένων καταλόγου, πρέπει να γίνουν διαπραγματεύσεις, συμφωνίες και ρυθμίσεις που αφορούν τη σύνταξη, το περιεχόμενο και την ποιότητα των δεδομένων. Σε αυτό το πλαίσιο, επειδή η σχέση μεταξύ παραλήπτη και αποστολέα είναι σαφής, το πρόβλημα είναι αρκετά μικρότερο. Όταν όμως πρόκειται για ενσωμάτωση δεδομένων Ιστού, η ενοποίηση της δομής προϊόντων έρχεται αντιμέτωπη με την περιορισμένη γνώση των τοπικών δομών προϊόντων, την ύπαρξη μεγάλου αριθμού τοπικών δομών-σχημάτων και πιθανές συχνές αλλαγές στα τοπικά σχήματα. Σε Peer to Peer (P2P) ταιριάσματα ακόμη και το ποιό είναι το επερώτημα και ποιά τα δεδομένα είναι θέμα οπτικής, και οι B2B αγορές μπορεί να είναι και πάροχοι και καταναλωτές των δεδομένων. Όσον αφορά στην ενοποίηση δεδομένων καταλόγου, είναι πολύ πιθανό ότι μέρη τουλάχιστον των μελλοντικών καταλόγων θα συναρμολογούνται από πράκτορες που θα εμπλουτίζουν τα δεδομένα εισόδου με πρόσθετες πληροφορίες. Αυτό δημιουργεί πρόσθετες απαιτήσεις οι οποίες πρέπει να αντιμετωπιστούν κατάλληλα εκ των προτέρων, ώστε να αποφευχθούν αντικανονικές λειτουργίες συμπερασμού με δυνητικά αρνητικές συνέπειες για τον αρχικό εκδότη των δεδομένων. Είναι προς το συμφέρον κάθε οντότητας - επιχείρησης που δημοσιεύσει στοιχεία να συμβάλλει στην ορθότητα των συμπερασμάτων που βασίζονται στα στοιχεία της, ώστε να αποφευχθούν οι δυσαρεστημένοι πελάτες, η εκτενής εξυπηρέτηση των πελατών ή παρόμοιες μειωνεκτικές καταστάσεις. Αυτή η κατάσταση είναι διαφορετική από την σημερινή στον Ιστό, όπου συνήθως δεν είναι αρνητικό αν οι μηχανές αναζήτησης αναφέρουν ψευδώς την σελίδα σου για ένα συγκεκριμένο όρο

αναζήτησης. Όσο όμως οι καταναλωτές των δεδομένων είναι μηχανές, τέτοιες ψευδώς θετικές καταστάσεις μπορεί να είναι τόσο αρνητικές όσο ένα ψευδώς αρνητικό αποτέλεσμα. Με λίγα λόγια, έγγραφα που σχετίζονται με προϊόντα και είναι δημοσιευμένα στον Σημασιολογικό Ιστό πρέπει να παρέχουν μια εκφραστικότητα που κάνει τα δεδομένα αναγνώσιμα από οποιαδήποτε μηχανή για μια πληθώρα σκοπών.



Σχήμα 2.5 : Ένα τυπικό σενάριο χρήσης δεδομένων προϊόντων στον Σημασιολογικό Ιστό.

2. *Κατανεμημένη Φύση* : Η πληροφορία για ένα προϊόν θα ανακτάται και θα συναρμολογείται από πολλά έγγραφα αποθηκευμένα σε πολλά διαφορετικά συστήματα. Στον Σημασιολογικό Ιστό μια συνηθισμένη κατάσταση θα είναι ότι (1) οι κατασκευαστές των προϊόντων δημοσιεύουν δεδομένα για τα προϊόντα τους σε έγγραφα του Παγκόσμιου Ιστού, (2) οι αντιπρόσωποι παρέχουν πρόσθετες πληροφορίες και τιμές, και (3) τρίτες οντότητες παρέχουν πρόσθετα στοιχεία, πληροφορίες, γνώση ή υπηρεσίες (π.χ. συνιστώμενα προϊόντα για ένα συγκεκριμένο σκοπό). Αυτό σημαίνει ότι θα έχουμε να αντιμετωπίσουμε προτάσεις που θα είναι αποθηκευμένες σε πολλαπλές πηγές από διαφορετικές επιχειρηματικές οντότητες, για επιμέρους σκοπούς

και καθοδηγούμενα από διαφορετικά κίνητρα. Είναι πιθανό δυο προτάσεις να αντιφάσκουν μεταξύ τους ή οι εκδότες να ανήκουν σε διαφορετικές σημασιολογικές κοινότητες. Αυτό το κατανεμημένο περιβάλλον σημαίνει ότι οποιαδήποτε συλλογιστική (reasoning) πρέπει να (1) χειρίζεται σωστά ελλιπείς πληροφορίες, (2) διακρίνει πληροφορίες που λείπουν από τις πληροφορίες άρνησης, και (3) εντοπίζει ακριβή ταιριάσματα, πιθανά ταιριάσματα και μερικά ταιριάσματα. Η απουσία ενός χαρακτηριστικού σε μια περιγραφή θα πρέπει να αντιμετωπίζεται ως χαρακτηριστικό που θα μπορούσε είτε να βελτιωθεί αργότερα ή να αγνοηθεί (αν είναι άσχετο). Οι ελλιπείς πληροφορίες είναι πολύ συνηθισμένη κατάσταση στον Σημασιολογικό Ιστό. Η βελτίωση μπορεί να επιτευχθεί με δυο τρόπους: Είτε χρειαζόμαστε ένα πρωτόκολλο για την αίτηση πρόσθετων χαρακτηριστικών, π.χ. μια υπηρεσία Ιστού (Web Service), ή θα πρέπει να αποκτήσουμε το κομμάτι της πληροφορίας που λείπει με χρήση συλλογιστικών και/ή χρησιμοποιώντας δεδομένα από πρόσθετες πηγές. Εκτός από αυτό, θα πρέπει να διασφαλίζεται ότι οι πιο συγκεκριμένες περιγραφές δεν θα θεωρούνται κατώτερες από τις τελείως γενικές ή τις απλώς λιγότερο συγκεκριμένες περιγραφές. Διαφορετικά, θα συμβεί το παράδοξο όπου όσο λιγότερες πληροφορίες παρέχεις τόσο πιο συχνά θα θεωρείσαι πιθανό ταίριασμα – αγνοώντας την αξία της συγκεκριμένης πληροφορίας για μια πιθανή επιχειρηματική συνεργασία. Είναι σημαντικό να σημειωθεί ότι οι εταιρίες έχουν μια εσωτερική αναπαράσταση του προϊόντος, συνήθως αποθηκευμένη στο ERP σύστημά τους, και ότι αυτό θα είναι συνήθως η πιο αξιόπιστη πηγή για τα δεδομένα των προϊόντων. Έτσι, η γεφύρωση του χάσματος μεταξύ της εσωτερικής ERP αναπαράστασης και των δεδομένων που δημοσιεύονται στον Σημασιολογικό Ιστό θα είναι μια βασική αποστολή για τις επιχειρήσεις.

3. *Η Δυναμική και η Μεταβλητότητα των Εννοιών* : Οι έννοιες αλλάζουν με την πάροδο του χρόνου και νέες έννοιες προκύπτουν. Οι σημασιολογικές αναπαραστάσεις προϊόντων είναι σχετικά σταθερές σε σύγκριση με τα δεδομένα συγκεκριμένου προμηθευτή. Ωστόσο, εξακολουθεί να υπάρχει

σημαντική δυναμική στον τομέα των προϊόντων, καθώς νέες κατηγορίες προϊόντων προκύπτουν και παλιές κατηγορίες προϊόντων γίνονται παράταιρες. Έτσι, η ορολογία προϊόντων που χρειάζονται αγοραστές και πωλητές είναι δυναμική. Με άλλα λόγια, οι αναπαραστάσεις που σχετίζονται με προϊόντα μπορεί να απαιτούν αλλαγές σε βάθος χρόνου. Αυτό δημιουργεί την ανάγκη για διαρκή συντήρηση. Βασικά, η δυναμική και η αστάθεια των εννοιών έχει άμεση σχέση με την εξειδίκευσή τους. Όσο μεγαλύτερη είναι η συγκεκριμενοποίηση των εννοιών, τόσο πιο δυναμικό είναι το λεξιλόγιο. Μια σημασιολογική αναπαράσταση δυο εννοιών, π.χ. “Πράγμα” και “Σκέψη” θα λειτουργήσουν χωρίς αλλαγή απείρως, και μια υπερβολικά πολύπλοκη αναπαράσταση που περιλαμβάνει στιγμιότυπα εννοιών για κάθε φαινόμενο στη γη σε κάθε δεδομένη χρονική στιγμή (“Αθήνα την 1η Απριλίου 2004", “ Αθήνα την 2η Απριλίου 2004”) θα ήταν το ακραίο αντίθετο με άπειρη δυναμική. Επειδή το κατώτερο επίπεδο σημασιολογικής ακρίβειας περιορίζει την ποιότητα των πιθανών αντιστοιχίσεων, μια σημασιολογική αναπαράσταση πρέπει να παρέχει έναν πολύ λεπτομερή διαχωρισμό. Έτσι, μια κατάλληλη αναπαράσταση των προϊόντων στον Σημασιολογικό Ιστό θα απαιτήσει μια υψηλού βαθμού σημασιολογική λεπτομέρεια, η οποία υπερβαίνει κατά πολύ τη διασπορά των σημερινών περιγραφικών γλωσσών.

2.2 Οντολογίες Σημασιολογικού Ιστού

2.2.1 Περιγραφή Οντολογιών και OWL

Βασικό συστατικό στοιχείο του Σημασιολογικού Ιστού αποτελούν οι Οντολογίες. Υπάρχουν διάφοροι ορισμοί για την έννοια της οντολογίας. Ο όρος είναι δανεισμένος από την Φιλοσοφία, όπου οντολογία είναι η περιγραφή των διαφόρων ειδών οντοτήτων στον κόσμο και του τρόπου που αυτές συσχετίζονται μεταξύ τους. Σύμφωνα με τον T.Gruber [25], μια οντολογία είναι μια τυπική, ρητή προδιαγραφή μιας κοινής εννοιολογικής θεώρησης ενός φαινομένου. Η εννοιολογική θεώρηση αφορά στα αντικείμενα, στις έννοιες και στις άλλες οντότητες που υποτίθεται ότι υπάρχουν σε κάποια περιοχή ενδιαφέροντος και τις

συσχετίσεις που υπάρχουν μεταξύ τους. Η εννοιολογική θεώρηση είναι μια αφηρημένη, απλοποιημένη όψη του κόσμου που θέλουμε να αναπαραστήσουμε.

Οι οντολογίες τυπικά εκφράζονται σε κάποια λογική γλώσσα (π.χ. λογική πρώτης τάξης). Στον Σημασιολογικό Ιστό οι οντολογίες συνήθως εκφράζονται σε γλώσσες οντολογιών όπως η RDFS και η OWL που βασίζονται σε αντικείμενα (individuals/objects), έννοιες (concepts/classes), συσχετίσεις (relations) και αξιώματα (axioms).

Η γλώσσα οντολογιών OWL 2 (Web Ontology Language) [4] είναι η πρόταση του W3C από τον Οκτώβριο 2009 για την δημιουργία οντολογιών Ιστού. Επεκτείνει τις RDF, RDFS γλώσσες, παρέχοντας μεγαλύτερες εκφραστικές δυνατότητες. Βασίζεται σε καλά ορισμένες έννοιες και προέρχεται από την λογική Description Logics (DLs). Χρησιμοποιείται για την αναπαράσταση γνώσης για κάποιο πράγμα, ομάδες πραγμάτων και σχέσεις μεταξύ τους.

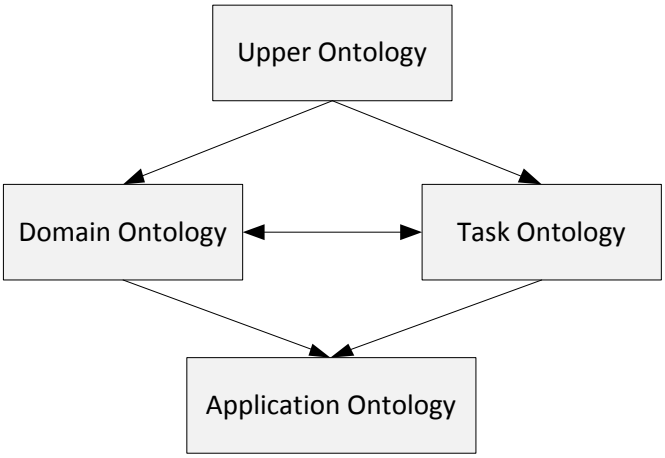
Τα πράγματα ή τα αντικείμενα για τα οποία αναπαρίσταται η γνώση ονομάζονται *στιγμιότυπα* (*instances*) (π.χ. ο Γιάννης, η Μαίρη). Οι ομάδες των πραγμάτων ονομάζονται *κλάσεις* (*classes*) (π.χ. Γυναίκα). Οι σχέσεις μεταξύ πραγμάτων (π.χ. παντρεμένος) ονομάζονται *ιδιότητες* (*properties*). Τα στιγμιότυπα, οι κλάσεις και οι ιδιότητες ονομάζονται *οντότητες* (*entities*).

Όπως συμβαίνει με τις Description Logics, οι οντότητες μπορούν να συνδυαστούν με τη χρήση *κατασκευαστών* (*constructors*) και να δημιουργηθούν πιο σύνθετες περιγραφές που ονομάζονται *εκφράσεις* (*expressions*). (π.χ. Άνδρας ⊔ Γυναίκα, όπου το σύμβολο ⊔ σημαίνει ότι δημιουργείται μια έκφραση που σημαίνει Άνδρας ή Γυναίκα). Για την αναπαράσταση γνώσης στην OWL, όπως σε κάθε γλώσσα αναπαράστασης γνώσης, κάνουμε δηλώσεις που ονομάζονται *αξιώματα* (*axioms*).

Οι οντότητες, οι εκφράσεις και τα αξιώματα αποτελούν το *λογικό μέρος* (*logical part*) μιας οντολογίας OWL, γιατί μπορεί να τους δοθεί ακριβή σημασιολογία και να εκτελέσουμε πρακτικές συμπερασμού πάνω τους. Εκτός από το λογικό μέρος υπάρχει και ο *σχολιασμός* (*annotation*) (παράδειγμα, ο σχολιασμός μιας κλάσης με έναν πιο περιγραφικό τίτλο από ό,τι είναι το όνομά της) που εφαρμόζεται στο

λογικό μέρος (δηλαδή στις οντότητες, αξιώματα, οντολογίες) αλλά δεν έχει καμιά επίδραση πάνω της.

Μια από τις χρησιμότητες των οντολογιών είναι η επαναχρησιμοποίηση γνώσης που επιτυγχάνεται εάν μια οντολογία που δημιουργείται για ένα πεδίο είναι (σε κάποιο βαθμό τουλάχιστον) επαναχρησιμοποιήσιμη σε άλλες εφαρμογές. Για να απλοποιηθεί τόσο η ανάπτυξη οντολογιών όσο και η επαναχρησιμοποίησή τους, είναι χρήσιμος ο σχεδιασμός τους σε τμήματα. Αυτός ο σχεδιασμός χρησιμοποιεί κληρονομικότητα οντολογιών, οι υψηλότερες στην ιεραρχία οντολογίες περιγράφουν γενική γνώση και οι οντολογίες εφαρμογών περιγράφουν γνώση για συγκεκριμένη εφαρμογή (Σχήμα 2.6).



Σχήμα 2.6 : Τμηματοποίηση οντολογιών.

Τα τμήματα που αποτελούν αυτή την ιεραρχία οντολογιών είναι :

- *Upper Ontology* : Γενική, υψηλότερου επιπέδου οντολογία που περιγράφει γενική γνώση όπως τι είναι χώρος και τι χρόνος.
- *Domain Ontology* : Οντολογία που περιγράφει ένα συγκεκριμένο πεδίο ενδιαφέροντος, όπως ο ιατρικός τομέας, ή ο τομέας της ηλεκτρονικής μηχανικής ή μικρότερους τομείς, όπως είναι οι προσωπικοί υπολογιστές.
- *Task Ontology* : Οντολογία κατάλληλη για συγκεκριμένη εργασία, όπως είναι η συναρμολόγηση εξαρτημάτων.

- *Application Ontology* : Οντολογία που αναπτύσσεται για συγκεκριμένη εφαρμογή όπως είναι η συναρμολόγηση προσωπικού υπολογιστή, που θα μπορούσε να είναι συνένεωση μιας domain οντολογίας για προσωπικούς υπολογιστές και μιας task οντολογίας για συναρμολόγηση εξαρτημάτων.

Στην εργασία αυτή χρησιμοποιείται οντολογία προϊόντων που είναι γραμμένη σε γλώσσα OWL και κατατάσσεται στο τμήμα των domain οντολογιών.

2.2.2 Η Μάθηση Οντολογίας

Ο Σημασιολογικός Ιστός στηρίζεται σε μεγάλο βαθμό στις επίσημες οντολογίες που δομούν τα υποκείμενα δεδομένα [17]. Γι αυτό η επιτυχία του Σημασιολογικού Ιστού εξαρτάται κυρίως από την διάδοση των οντολογιών, πράγμα που απαιτεί γρήγορη και εύκολη μηχανική (engineering) οντολογιών για αποφυγή συμφόρησης στην απόκτηση γνώσης.

Η μάθηση οντολογίας διευκολύνει την κατασκευή οντολογιών από τον μηχανικό οντολογιών (ontology engineer). Οι οντολογίες είναι σχήματα (μετα)δεδομένων που παρέχουν ένα ελεγχόμενο λεξιλόγιο εννοιών, το καθένα με μια ρητά ορισμένη και επεξεργάσιμη από τις μηχανές σημασιολογία. Με τον καθορισμό κοινών και διαμοιραζόμενων εννοιών σε τομείς, οι οντολογίες βοηθούν και τους ανθρώπους και τις μηχανές να επικοινωνούν συνοπτικά, υποστηρίζοντας την ανταλλαγή σημασιολογίας και όχι μόνο σύνταξης. Ως εκ τούτου, η φθηνή και γρήγορη κατασκευή οντολογιών είναι κρίσιμη για την επιτυχία και τον ανάπτυξη του Σημασιολογικού Ιστού.

Η απόκτηση γνώσεων σε ένα τομέα για την κατασκευή οντολογιών απαιτεί πολύ χρόνο και πολλούς πόρους. Με αυτή την έννοια μπορούμε να ορίσουμε την μάθηση οντολογίας ως το σύνολο των μεθόδων και των τεχνικών που χρησιμοποιούνται για την κατασκευή μιας οντολογίας από το μηδέν, τον εμπλουτισμό ή την προσαρμογή μιας υπάρχουσας οντολογίας με ένα ημι-αυτόματο τρόπο χρησιμοποιώντας διάφορες πηγές [16]. Άλλοι όροι έχουν χρησιμοποιηθεί επίσης για να δηλώσουν την ημιαυτόματη κατασκευή οντολογιών

όπως παραγωγή οντολογίας (ontology generation), εξόρυξη οντολογίας (ontology mining), εξαγωγή οντολογίας (ontology extraction) κ.ά. Διάφορες τεχνικές υπάρχουν για την μερική αυτοματοποίηση της διαδικασίας απόκτησης γνώσης, οι οποίες εξαρτώνται από τύπο της πηγής δεδομένων, για παράδειγμα όταν η πηγή είναι απλό κείμενο τότε χρησιμοποιούνται τεχνικές ανάλυσης φυσικής γλώσσας και μηχανική μάθηση. Αναλυτικά οι προσεγγίσεις μάθησης οντολογίας περιγράφονται στην επόμενη υποενότητα. Η χρησιμότητα της μάθησης οντολογίας είναι μεγάλη διότι υπόσχεται να ξεπεράσει κάποια προβλήματα της μηχανικής οντολογιών (ontology engineering), όπως είναι η απαίτηση για γρήγορη ανάπτυξη οντολογιών, με σχετικά εύκολο τρόπο και παράλληλα η οντολογία που θα προκύψει να είναι έγκυρη.

Η μάθηση οντολογίας (ontology learning) περιλαμβάνει τις μεθόδους και τις τεχνικές που χρησιμοποιούνται για 1) δημιουργία οντολογίας από το μηδέν ή 2) εμπλουτισμό ή προσαρμογή μιας υπάρχουσας οντολογίας με ημι-αυτόματο τρόπο και τη χρήση διαφόρων πηγών πληροφοριών [12]. Επομένως, η μάθηση οντολογίας δεν είναι μόνο η διαδικασία της αρχικής εξαγωγής. Στο πλαίσιο αυτό ο εμπλουτισμός μιας οντολογίας με νέες κλάσεις και στιγμιότυπα θεωρείται επίσης μάθηση οντολογίας και αυτό το κομμάτι αποτελεί αντικείμενο αυτής της εργασίας, που σκοπό έχει τον εμπλουτισμό μιας υπάρχουσας οντολογίας με προϊόντα από πίνακες και html λίστες σελίδων Ιστού.

Παρά το γεγονός ότι δεν υπάρχει ένας κοινά αποδεκτός ορισμός για το έργο της συμπλήρωσης οντολογίας (ontology population), μια χρήσιμη προσέγγιση έχει προταθεί [46], ως εξαγωγή πληροφορίας προσανατολισμένη σε οντολογία (Ontology Driven Information Extraction), όπου στη θέση του προτύπου που πρέπει να πληρωθεί, είναι η εξόρυξη και η ταξινόμηση των στιγμιοτύπων των εννοιών και των σχέσων που ορίζονται σε μια οντολογία. Η μάθηση οντολογιών (ontology learning) είναι διαφορετική από το ontology population ,καθώς δεν αφορά μόνο την δημιουργία στιγμιοτύπων, αλλά είναι η διαδικασία που υποτίθεται ότι αποκτούνται νέες έννοιες και σχέσεις με συνέπεια την αλλαγή του ορισμού της ίδιας της οντολογίας.

2.2.3 Προσεγγίσεις Μάθησης Οντολογίας

Η μάθηση οντολογίας μπορεί να διακριθεί σε διάφορες προσεγγίσεις ανάλογα με τον τύπο εισόδου-πηγή δεδομένων που χρησιμοποιείται στην μάθηση [16]. Με αυτή την έννοια, προτείνεται η εξής ταξινόμηση: μάθηση οντολογίας από κείμενο, από λεξικό (dictionary), από βάση γνώσης, από ημι-δομημένα σχήματα και από σχεσιακά σχήματα. Η μάθηση οντολογίας θέτει έναν αριθμό ερευνητικών δραστηριοτήτων που επικεντρώνονται σε διαφορετικούς τύπους εισόδου, αλλά μοιράζονται το στόχο τους για κοινή εννοιολογική μοντελοποίηση. Είναι ένα περίπλοκο διεπιστημονικό πεδίο που χρησιμοποιεί τη γνώση της επεξεργασίας φυσικής γλώσσας, της εξόρυξης δεδομένων και δικτύων, της μηχανικής μάθησης και της αναπαράστασης γνώσης.

Οι μέθοδοι *μάθησης οντολογίας από κείμενα* αποτελούνται από εξαγωγή οντολογιών εφαρμόζοντας τεχνικές ανάλυσης φυσικής γλώσσας σε κείμενα. Οι πιο γνωστές προσεγγίσεις αυτής της ομάδας είναι:

- *Εξόρυξη που βασίζεται σε πρότυπα (pattern-based extraction)*. Μια σχέση αναγνωρίζεται όταν μια ακολουθία λέξεων στο κείμενο ταιριάζει με ένα μοτίβο. Για παράδειγμα, ένα μοτίβο μπορεί να καθιερώσει ότι αν εντοπιστεί μια ακολουθία n ονομάτων, τότε τα n-1 πρώτα ονόματα είναι hyponyms του n-οστού.
- *Κανόνες συσχέτισης (association rules)*. Είχαν οριστεί αρχικά στο πεδίο των βάσεων δεδομένων ως εξής: “Δεδομένου ενός συνόλου συναλλαγών, όπου κάθε συναλλαγή είναι ένα σύνολο από literals (τα ονομαζόμενα στοιχεία), ένας κανόνας συσχέτισης είναι μια έκφραση της μορφής X συνεπάγεται Y, είναι σύνολα στοιχείων. Η διαισθητική σημασία ενός τέτοιου κανόνα είναι ότι οι συναλλαγές της βάσης που περιλαμβάνουν το X τείνουν να περιλαμβάνουν και το Y“. Οι κανόνες συσχέτισης χρησιμοποιούνται σε διαδικασίες εξόρυξης δεδομένων για την ανακάλυψη πληροφορίας που είναι αποθηκευμένη στις βάσεις εάν από πριν έχουμε κάποια ιδέα για το τι ψάχνουμε. Η μέθοδος των κανόνων συσχέτισης για μάθηση οντολογίας έχει χρησιμοποιηθεί για την

ανακάλυψη σχέσεων μεταξύ εννοιών εκτός των ιεραρχικών, χρησιμοποιώντας σαν γνωστικό υπόβαθρο μια ιεραρχία εννοιών.

- *Εννοιολογική ομαδοποίηση (conceptual clustering).* Οι έννοιες ομαδοποιούνται σύμφωνα με την σημασιολογική απόσταση μεταξύ τους για να δημιουργηθούν ιεραρχίες. Για να υπολογιστεί η σημασιολογική απόσταση μεταξύ δυο εννοιών ο τύπος (formulae) μπορεί να εξαρτάται από διαφορετικούς παράγοντες που πρέπει να δίνεται σε αυτές τις μεθόδους.
- *Κλάδεμα οντολογίας (ontology pruning).* Ο σκοπός του κλαδέματος οντολογίας είναι να χτιστεί μια οντολογία, σε κάποιο πεδίο ενδιαφέροντος, που θα βασίζεται σε διαφορετικές ετερογενείς πηγές. Περιλαμβάνει τα ακόλουθα βήματα. Πρώτα, μια γενική οντολογία πυρήνα χρησιμοποιείται σαν δομή υψηλοτέρου επιπέδου. Δεύτερο, χρησιμοποιείται ένα λεξικό που περιέχει σημαντικούς όρους του πεδίου εφαρμογής σε φυσική γλώσσα ώστε να αποκτηθούν οι έννοιες του πεδίου. Αυτές οι έννοιες ταξινομούνται στην γενική οντολογία πυρήνα. Τρίτο, κείμενα συγκεκριμένα του πεδίου εφαρμογής χρησιμοποιούνται για να αφαιρεθούν έννοιες που δεν ήταν συγκεκριμένες για το πεδίο. Η αφαίρεση εννοιών ακολουθεί τον ευριστικό κανόνα ότι έννοιες που είναι συγκεκριμένα για το πεδίο πρέπει να είναι πιο συχνές σε μια συλλογή κειμένων συγκεκριμένου πεδίου από ότι σε γενικά κείμενα.
- *Μάθηση εννοιών (concept learning):* Μια δεδομένη ταξινομία ενημερώνεται σταδιακά όσο αποκτούνται νέες έννοιες από κείμενα του πραγματικού κόσμου.

Η *μάθηση οντολογίας από λεξικό* βασίζει την απόδοσή της στην χρήση ενός λεξικού αναγνώσιμου από μηχανές, ώστε να εξάγει σχετικές έννοιες και συσχετίσεις μεταξύ τους.

Η *μάθηση οντολογίας από βάση γνώσης* αποσκοπεί στην μάθηση οντολογίας χρησιμοποιώντας ως πηγή υπάρχουσες βάσεις γνώσης.

Η *μάθηση οντολογίας από ημιδομημένα δεδομένα* αναζητά την απόσπαση οντολογίας από πηγές που έχουν οποιαδήποτε προκαθορισμένη δομή, όπως είναι τα XML σχήματα.

Η *μάθηση οντολογίας από σχεσιακά σχήματα* σκοπεύει στην μάθηση μιας οντολογίας με την εξαγωγή σχετικών εννοιών και συσχετίσεων από γνώση που βρίσκεται στις βάσεις δεδομένων.

ΚΕΦΑΛΑΙΟ 3

ΕΙΚΟΝΙΚΕΣ ΑΓΟΡΕΣ – ΟΝΤΟΛΟΓΙΕΣ

3.1 Εικονικές Αγορές

3.1.1 Περιγραφή Εικονικών Αγορών

Οι εικονικές αγορές (electronic marketplaces) που βασίζονται στο Internet γίνονται ολοένα και πιο δημοφιλείς. Εμφανίζονται σε διάφορες βιομηχανίες, υποστηρίζοντας την ανταλλαγή αγαθών και υπηρεσιών διαφόρων ειδών, με διαφορετικούς τύπους φορέων και ακολουθούν διαφορετικές αρχιτεκτονικές αρχές. Οι περισσότεροι ερευνητές πιστεύουν ότι οι εικονικές αγορές έχουν έρθει για να κυριαχήσουν στο τοπίο του ηλεκτρονικού επιχειρείν (e-business) [6].

Έχουν δοθεί διάφοροι ορισμοί για την έννοια της εικονικής αγοράς. Μια αγορά, ως ένας ιστορικά εξελίξιμος θεσμός επιτρέπει στους πελάτες και τους προμηθευτές να συναντηθούν σε ορισμένο τόπο και χρονικό διάστημα, προκειμένου να επικοινωνήσουν και να ανακοινώσουν τις προθέσεις αγοράς ή πώλησης, που τελικά ταιριάζουν και μπορούν να παραγματοποιηθούν [6]. Σήμερα, εξακολουθεί να συμβαίνει το ίδιο, αλλά η αγορά έχει κατά καιρούς ξανα-μοντελοποιηθεί λόγω της εξέλιξης των μέσων. Έτσι, λόγω της εξέλιξης των σύγχρονων τεχνολογιών και τηλεπικοινωνιών, οι περιορισμοί χρόνου και χώρου έχουν αποδυναμωθεί και ο κυβερνοχώρος έχει γίνει το νέο σημείο συνάντησης. Το ιδιαίτερο χαρακτηριστικό μιας εικονικής αγοράς είναι ότι φέρνει *πολλαπλούς* αγοραστές και πωλητές μαζί (με την «εικονική» έννοια) σε ένα κεντρικό χώρο της αγοράς. Μια περιγραφή των εικονικών αγορών ορίζει μια εικονική αγορά (e-marketplace) ως μια εικονική online αγορά όπου οργανώσεις εγγράφονται ως αγοραστές ή πωλητές για τη διεξαγωγή Business to Business (B2B) ηλεκτρονικού εμπορίου μέσω του διαδικτύου [7]. Υπάρχουν πολλοί τύποι εικονικών αγορών βασιζόμενοι σε μια σειρά από επιχειρησιακά μοντέλα. Μπορούν να λειτουργήσουν μέσω μιας ανεξάρτητης τρίτης οντότητας ή να

διοικούνται από κάποια μορφή κοινοπραξίας που έχει συσταθεί για να εξυπηρετήσει ένα συγκεκριμένο τομέα ή αγορά.

Τα συστατικά στοιχεία μιας εικονικής αγοράς είναι οι αγοραστές, οι πωλητές και τα προϊόντα ή οι υπηρεσίες. Ψηφιακά προϊόντα είναι τα αγαθά που μπορούν να μετατραπούν σε ψηφιακή μορφή και να σταλούν μέσω του Διαδικτύου. Άλλα συστατικά στοιχεία μιας εικονικής αγοράς είναι η υποδομή της, το front-end, το back-end και οι μεσάζοντες, που είναι τρίτες οντότητες που λειτουργούν μεταξύ αγοραστών και πωλητών. Επίσης μπορεί να υπάρχουν άλλοι επιχειρηματικοί εταίροι και υποστηρίζουσες υπηρεσίες.

Οι εικονικές αγορές (electronic markets/e-markets) [5] είναι το θεμέλιο του ηλεκτρονικού εμπορίου. Δυνητικά, ενοποιούν τη διαφήμιση, την παραγγελία προϊόντων, την διανομή ψηφιακών προϊόντων και τα συστήματα πληρωμής. Μια εικονική αγορά είναι ένα διεπιχειρησιακό πληροφοριακό σύστημα που επιτρέπει στους συμμετέχοντες αγοραστές και πωλητές να ανταλλάξουν πληροφορίες σχετικά με τιμές και προσφερόμενα προϊόντα. Η εταιρεία που λειτουργεί αυτό το σύστημα αναφέρεται ως διαμεσολαβητής, που μπορεί να είναι κάποιος που συμμετέχει στην αγορά, δηλαδή αγοραστής ή πωλητής ή μια ανεξάρτητη τρίτη οντότητα ή μια κοινοπραξία πολλών εταιρειών. Οι εικονικές αγορές παρέχουν μια ηλεκτρονική ή online μέθοδο για την διευκόλυνση συναλλαγών μεταξύ αγοραστών και πωλητών και παρέχει υποστήριξη σε όλα τα βήματα της διαδικασίας ολοκλήρωσης παραγγελίας. Το μοντέλο επιχειρησιακής διαδικασίας από την πλευρά του καταναλωτή αποτελείται από δραστηριότητες που μπορούν να ομαδοποιηθούν σε τρεις φάσεις: (1) αποφασιστικότητα πριν την αγορά, (2) κατανάλωση αγορασθέντος και (3) αλληλεπίδραση μετά την αγορά. Κάθε μια από αυτές τις φάσεις μπορεί να υποστηριχθεί ηλεκτρονικά σε μια ολοκληρωμένη εικονική αγορά, αλλά οι εικονικές αγορές σήμερα γενικά υποστηρίζουν μόνο τις δραστηριότητες αποφασιστικότητας πριν την αγορά, αν και κινούνται προς την κατεύθυνση της κατανάλωσης αγορασθέντος.

Για να συνοψίσουμε, μια εικονική αγορά θα μπορούσε να οριστεί [6]:

- Θεσμικά, ως ένα μέσο.
  - που αναθέτει διαφορετικούς ρόλους μέσα σε μια κοινότητα, κατά κύριο λόγο αγοραστές και προμηθευτές, αλλά και άλλους ρόλους, όπως πάροχοι υπηρεσιών εφοδιασμού, τράπεζες και άλλους διαμεσολαβητές.
  - που διευκολύνουν την ανταλλαγή πληροφοριών, αγαθών, υπηρεσιών και πληρωμών.
  - που παρέχει μια υποδομή – καθορίζει τα πρωτόκολλα και τις διαδικασίες που κανονίζουν τις αλληλεπιδράσεις μέσα στην κοινότητα, καθώς επίσης παρέχουν μια κοινή γλώσσα.
- Κοινωνικά, ως μια κοινότητα που αποτελείται από αγοραστές, πωλητές, κτλ.
  - η οποία θα μπορούσε να περιγραφεί από μια ορισμένη κατάσταση, που περιλαμβάνει τη γνώση των συμμετεχώντων, την πρόθεση, τις συμβάσεις (στοιχεία του ενεργητικού και παθητικού) και τα εμπορεύματα, σε ένα ορισμένο χρονικό διάστημα.
  - με τους ρόλους που περιλαμβάνουν, τα δικαιώματα και τις υποχρεώσεις
  - που προτίθεται να χρησιμοποιήσει την αγορά συναλλάγματος – ή διαδικασίες επικοινωνίας – ώστε να αλλάξουν την κατάστασή τους, σύμφωνα για τις προθέσεις τους.

Οι εικονικές αγορές μπορούν να μελετηθούν και από άλλες ακοπιές όπως για παράδειγμα την οικονομική ή την νομική σκοπιά.

Οι λειτουργίες μιας εικονικής αγοράς συνοψίζονται στις εξής :

- *Ταίριασμα αγοραστών και πωλητών* :
  - Προσδιορισμός προϊόντων για προσφορά. Περιγράφονται οι ιδιότητες των προϊόντων που προσφέρονται από τους πωλητές και γίνεται συνάθροιση διαφορετικών προϊόντων.
  - Αναζήτηση (αγοραστών για πωλητές και πωλητών για αγοραστές). Ανταλλάσονται πληροφορίες για προϊόντα και τιμές. Γίνεται οργάνωση

προσφορών και παζαριών. Επιπλέον γίνεται ταίριασμα των προσφορών των πωλητών με τις επιθυμίες των αγοραστών.

- Ανακάλυψη τιμών. Περιλαμβάνει την διαδικασία και τα αποτελέσματα στον καθορισμό τιμών και δίνει την δυνατότητα για συγκρίσεις τιμών.
- Άλλες λειτουργίες, όπως είναι η παροχή ευκαιριών πωλήσεων.

➢ *Διευκόλυνση συναλλαγών* :

- Υλικοτεχνική υποστήριξη (logistics), που περιλαμβάνει τις πληροφορίες αποστολής αγαθών ή υπηρεσιών στους αγοραστές.
- Διακανονισμός για την μεταφορά των πληρωμών στους πωλητές.
- Εμπιστοσύνη. Αυτό επιτυγχάνεται με τη χρήση πιστωτικών συστημάτων (credit systems) και καλής φήμης της εικονικής αγοράς. Επίσης οι εικονικές αγορές αξιολογούνται μέσω των οργανισμών αξιολόγησης και ειδικών online οργανισμών διαχείρισης εμπιστοσύνης (trust agencies).
- Επικοινωνία. Ανάρτηση αιτημάτων αγοραστών.

➢ *Θεσμική Υποδομή* :

- Νομική σκοπιά. Περιλαμβάνει τον εμπορικό κώδικα, τις συμβάσεις δικαίου, την επίλυση διαφορών και την προστασία πνευματικών διακιωμάτων που πρέπει να τηρούν οι εικονικές αγορές. Επίσης εδώ συμπεριλαμβάνονται και οι νόμοι που ισχύουν για εισαγωγές και εξαγωγές.
- Ρυθμιστικοί κανονισμοί. Είναι οι νόμοι και κανόνες που ισχύουν σε μια εικονική αγορά, η παρακολούθηση εφαρμογής τους και η επιβολή τους.
- Ανακάλυψη εικονικής αγοράς. Περιλαμβάνει πληροφορίες για την εικονική αγορά, όπως για παράδειγμα για τον ανταγωνισμό που υπάρχει ή τους κρατικούς κανονισμούς στις οποίες υπόκειται η (εικονική) αγορά.

Ένα πλήθος εικονικών αγορών είναι διαθέσιμο στους καταναλωτές για την αγορά προϊόντων από μουσικά CD μέχρι μηχανοκίνητα οχήματα. Μερικά προϊόντα και υπηρεσίες που υπάρχουν διαθέσιμα σε εικονικές αγορές αφορούν IT προϊόντα,

λουλούδια, ρουχισμό, μηχανοκίνητα, μουσική, βιβλία, ηλεκτρονικά περιοδικά, αεροπορικά εισητήρια, χρηματιστήριο αξιών κτλ. Η αυξανόμενη χρήση του Διαδικτύου έχει καταστήσει δυνατό να δημιουργηθούν νέες αγορές online. Ένα παράδειγμα είναι το eBay, ένας παγκόσμιος οίκος δημοπρασιών. Το Διαδίκτυο έχει επιτρέψει και άλλες αγορές να ευδοκιμήσουν με το να συνδέουν αγοραστές και πωλητές από απομακρυσμένες τοποθεσίες. Ο σχηματισμός online αγορών συνήθως συμβαίνει γρήγορα ως αποτέλεσμα κοινωνικών και οικονομικών τάσεων.

3.1.2 Κατηγορίες Εικονικών Αγορών

Στην βιβλιογραφία μπορεί κανείς να αναζητήσει και να βρει ένα σύνολο κατηγοριοποιήσεων των εικονικών αγορών. Στο [6], οι εικονικές αγορές διακρίνονται στις εξής κατηγορίες :

- αγορές *προσανατολισμένες στον αγοραστή (buyer-oriented)* : όπου η εικονική αγορά δημιουργείται από έναν συνεταιρισμό (consortium) αγοραστών, συνήθως προερχόμενων από τον ίδιο επιχειρηματικό κλάδο, οι οποίοι προμηθεύονται προϊόντα και υπηρεσίες μέσω του διαδικτύου. Μια εικονική αγορά προσανατολισμένη στον αγοραστή συνήθως διευθύνεται από μια κοινοπραξία αγοραστών, ώστε να δημιουργηθεί ένα αποτελεσματικό περιβάλλον αγοράς. Για κάποιον που ψάχνει να κάνει αγορές, η συμμετοχή σε αυτού του είδους την εικονική αγορά μπορεί να τον βοηθήσει να κάνει λιγότερα διοικητικά έξοδα και να επιτύχει την καλύτερη τιμή από τους προμηθευτές. Ως προμηθευτής μπορεί κανείς να χρησιμοποιήσει μια εικονική αγορά προσανατολισμένη στον αγοραστή για να διαφημίσει τον κατάλογό του σε μια ομάδα σχετικών πελατών που ψάχνουν να αγοράσουν.
- αγορές *προσανατολισμένες στον πωλητή (seller-oriented)* : όπου η εικονική αγορά δημιουργείται από έναν συνεταιρισμό προμηθευτών ή πωλητών, οι οποίοι πωλούν προϊόντα και υπηρεσίες μέσω του διαδικτύου. Επίσης γνωστή ως κατάλογος προμηθευτή, αυτή η αγορά έχει συσταθεί και λειτουργεί από μια σειρά προμηθευτών που επιδιώκουν να δημιουργηθεί ένα

αποτελεσματικό κανάλι πωλήσεων προς ένα μεγάλο αριθμό αγοραστών μέσω του διαδικτύου. Συνήθως η αναζήτηση αυτών των αγορών γίνεται με βάση το προϊόν ή την υπηρεσία που προσφέρεται. Οι κατάλογοι προμηθευτών οφελούν τους αγοραστές γιατί παρέχουν πληροφορίες για αγορές και περιοχές σχετικά με προμηθευτές που μπορεί να μην γνωρίζουν. Οι πωλητές μπορούν να χρησιμοποιούν αυτούς τους τύπους της αγοράς για να αυξήσουν την προβολή τους προς πιθανούς αγοραστές.

- *ουδέτερες εικονικές αγορές (neutral)* : όπου η εικονική αγορά δημιουργείται με σκοπό να προσελκύσει και αγοραστές και προμηθευτές, να τους φέρει σε επαφή ώστε να γίνουν εμπορικές συναλλαγές μεταξύ τους, χωρίς να δίνεται έμφαση σε κάποια από τις δυο πλευρές. Μια ανεξάρτητη εικονική αγορά είναι συνήθως μια B2B πλατφόρμα που λειτουργείται από μια τρίτη οντότητα η οποία είναι ανοικτή για τους αγοραστές ή πωλητές σε μια συγκεκριμένη βιομηχανία. Με την εγγραφή σε μια ανεξάρτητη εικονική αγορά, δίνεται πρόσβαση σε μικρές αγγελίες ή αιτήσεις για προσφορές ή προσφορές στον τομέα της βιομηχανίας. Συνήθως απαιτείται κάποια μορφή πληρωμής για τη συμμετοχή.

Μια άλλη κατηγοριοποίηση των εικονικών αγορών τις διακρίνει σε οριζόντιες και κάθετες:

- *οριζόντιες αγορές (horizontal)* : Μια οριζόντια εικονική αγορά συνδέει αγοραστές και πωλητές στους διάφορους κλάδους ή περιοχές. Μια οριζόντια εικονική αγορά μπορεί να χρησιμοποιηθεί για την αγορά έμμεσων προϊόντων, όπως εξοπλισμό γραφείου και γραφικής ύλης.
- *κάθετες εικονικές αγορές (vertical)* : Οι κάθετες εικονικές αγορές παρέχουν online πρόσβαση στις επιχειρήσεις κατακόρυφα προς τα πάνω και προς τα κάτω σε κάθε τμήμα ενός συγκεκριμένου τομέα της βιομηχανίας, όπως η αυτοκινητοβιομηχανία, τα χημικά, οι κατασκευές ή τα κλωστοϋφαντουργικά προϊόντα. Η αγορά ή πώληση σε μια κάθετη εικονική αγορά για τον τομέα μιας βιομηχανίας μπορεί να αυξήσει τη λειτουργική αποτελεσματικότητά της

και να βοηθήσει να μειωθεί το κόστος της εφοδιαστικής αλυσίδας, τα αποθέματα και ο χρόνος του κύκλου προμηθειών.

Σε άλλη κατηγοριοποίηση, οι εικονικές αγορές διακρίνονται σε :

- αγορές με *σταθερό μηχανισμό τιμολόγησης* : είναι αγορές, όπως οι ηλεκτρονικοί κατάλογοι όπου οι τιμές είναι σταθερές και τα ηλεκτρονικά καταστήματα δεν έχουν μηχανισμούς τιμολόγησης.
- αγορές με *μεταβλητό μηχανισμό τιμολόγησης* : είναι αγορές όπως οι δημοπρασίες ή οι ανταλλαγές που εφαρμόζονται δυναμικοί μηχανισμοί τιμολόγησης.

Τέλος, σε μια γενική κατηγοριοποίηση, οι εικονικές αγορές διακρίνονται και ως :

- *ανοιχτές αγορές (open)* : Μια ανοιχτή εικονική αγορά είναι αυτή που επιτρέπει την πρόσβαση σε όλους όσους δεν παραβιάζουν τους κανονισμούς της.
- *κλειστές αγορές (closed) :* έχουν πρόσβαση μόνο συγκεκριμένοι αγοραστές και πωλητές.

Αναφορικά με την διάκριση ανοιχτών και κλειστών εικονικών αγορών, υπάρχει η θεώρηση [35] για τον κανόνα reach/scope. Σημαίνει ότι μια αποτελεσματική εικονική αγορά με μεγάλη προσβάση (long reach), για παράδειγμα μια ανοιχτή δομή που επιτρέπει πολλούς συμμετέχοντες, θα έχει χαμηλό πεδίο εφαρμογής (low scope) (απλή λειτουργικότητα). Η επίτευξη πολύπλοκης λειτουργικότητας με πολλούς συμμετέχοντες είναι πολύ δύσκολο να επιτευχθεί. Εικονικές αγορές με κλειστή δομή και συνεπώς λίγους συμμετέχοντες (short reach) μπορεί να επιτύχει σύνθετη λειτουργικότητα (high scope).

Άλλες κατηγοριοποιήσεις των εικονικών αγορών, τις διακρίνουν βάσει της διαδικασίας αγοράς, «Τι» και «Πώς» αγοράζουν οι επιχειρήσεις (manufacturing vs operating inputs ή spot vs system sourcing), ή ανάλογα με τη στήριξη που δίνεται στις διάφορες φάσεις των συναλλαγών (την ανταλλαγή πληροφοριών, τη διαπραγμάτευση, την εκκαθάριση, μετά-την πώληση) ή τέλος, βάσει του μηχανισμού της αγοράς, άθροιση (aggregation) ή ταίριασμα (matching).

Οι τύποι εικονικών αγορών από τη σκοπιά της ιδιοκτησίας και της χρήσης προτύπων περιλαμβάνουν τις ιδιωτικές αγορές, τις περιφερειακές αγορές και τις κάθετες αγορές [8].

- Οι *ιδιωτικές αγορές* κατά κανόνα ανήκουν σε μια μεμονωμένη οργάνωση ή κοινοπραξίες οργανισμών, είτε σε αγοραστές ή πωλητές (για παράδειγμα ένας οργανισμός-αγοραστής μπορεί να έχει μια ιδιωτική αγορά που περιλαμβάνει μόνο τους δικούς της προμηθευτές).
- Οι *περιφερειακές αγορές* είναι παρόμοιες με τις ιδιωτικές αλλά ανήκουν σε μια ομάδα ή κοινοπραξίες αγοραστών που προέρχονται από μια γεωγραφική περιοχή. Οι περιφερειακές αγορές μπορεί να είναι μία μέθοδος που επιτρέπει στις τοπικές αρχές να συνεργαστούν με άλλες τοπικές αρχές για τις δημόσιες συμβάσεις. Οι περιφερειακές εικονικές αγορές επιτρέπουν διάφορες τοπικές αρχές να χρησιμοποιούν το ίδιο σύστημα ηλεκτρονικών προμηθειών ώστε να μοιράζονται τους προμηθευτές τους. Οφέλη μπορούν να επιτευχθούν μέσω συνεργατικής προμήθειας, διότι αξιοποιώντας ποσότητες αγορών μπορούν να διαπραγματεύονται καλύτερους όρους συμβολαίου.
- Οι *κάθετες αγορές* ειδικεύονται στην εμπορία των προϊόντων που σχετίζονται με έναν μόνο τομέα δραστηριότητας, όπως τα βιομηχανικά χημικά προϊόντα ή οι ιατρικές προμήθειες.

3.1.3 Οφέλη εικονικών αγορών

Η χρήση των εικονικών αγορών αυξάνεται διαρκώς διότι παρέχει πολλά οφέλη προς όλες τις συμμετέχουσες οντότητες. Τα επιχειρησιακά οφέλη μιας εικονικής αγοράς είναι οι μειωμένες χειρωνακτικές διαχειριστικές διεργασίες που αφορούν στην παραγγελία, η μείωση των λαθών, ο ταχύτερος κύκλος αγορών, η βελτιωμένη προβολή των αγορών, ο καλύτερος έλεγχος των δαπανών [8].

Τα βασικά κίνητρα των επιχειρήσεων για τη χρήση εικονικών αγορών είναι ο εξορθολογισμός των διαδικασιών, ώστε να αυξηθεί ο έλεγχος των δαπανών.

Επιπρόσθετα μια εικονική αγορά μπορεί να επιτρέψει στις οργανώσεις να έχουν καλύτερη ορατότητα του τι δαπανάται, σε ποια στοιχεία και από ποιον.

Τα οφέλη ανάμεσα στις διάφορες οντότητες της εικονικής αγοράς κατανέμονται ως ακολούθως.

- Γενικά επιχειρηματικά οφέλη.
    1. Υπάρχουν περισσότερες ευκαιρίες για τους προμηθευτές και τους αγοραστές για τη δημιουργία νέων εταιρικών συναλλαγών, είτε εντός της αλυσίδας εφοδιασμού τους ή σε ολόκληρη την αλυσίδα εφοδιασμού.
    2. Οι εικονικές αγορές μπορούν να προσφέρουν μεγαλύτερη διαφάνεια στη διαδικασία της αγοράς διότι η διαθεσιμότητα, οι τιμές και τα επίπεδα των αποθεμάτων είναι όλα προσβάσιμα σε ένα ανοικτό περιβάλλον.
    3. Οι χρονικοί περιορισμοί και τα προβλήματα με τις διαφορετικές ώρες λειτουργίας για το διεθνές εμπόριο δεν υπάρχουν πλέον, καθώς θα υπάρχει λειτουργία σε εικοσιτετράωρη βάση.
- Οφέλη για τον αγοραστή.
    1. Ενημερωμένες πληροφορίες για την τιμή και τη διαθεσιμότητα καθιστά ευκολότερο να εξασφαλίσει ο αγοραστής την καλύτερη διαπραγμάτευση.
    2. Οι εικονικές αγορές προσφέρουν ένα βολικό τρόπο για να συγκρίνουν τις τιμές και τα προϊόντα από μία μόνο πηγή και όχι να χάνουν χρόνο επικοινωνώντας με κάθε προμηθευτή.
    3. Οι καθιερωμένες εικονικές αγορές παρέχουν ένα επίπεδο εμπιστοσύνης για τον αγοραστή, γιατί έχουν να κάνουν αποκλειστικά και μόνο με προμηθευτές που είναι μέλη.
- Οφέλη για τον πωλητή.
    1. Είναι δυνατές οι τακτικές αιτήσεις για προσφορές τιμών από νέους και υπάρχοντες πελάτες.
    2. Παρέχει ένα επιπλέον κανάλι πωλήσεων ως αγορά για να πωλούν τα προϊόντα.

3. Οι εικονικές αγορές μπορούν να προσφέρουν μείωση του κόστους μάρκετινγκ, σε σύγκριση με άλλα κανάλια διάθεσης.
4. Η χρήση διεθνών εικονικών αγορών μπορεί να προσφέρει ευκαιρίες για πωλήσεις στο εξωτερικό για τις οποίες διαφορετικά δεν θα μπορούσε να γνωρίζει.

3.2 Οντολογίες Προϊόντων και Εικονικές Αγορές

Ένας ορισμός για το τι είναι *οντολογίες προϊόντων (product ontologies)* λέει ότι οντολογίες προϊόντων είναι τα πρότυπα περιεχομένου που μπορούν να χρησιμοποιηθούν για την αναπαράσταση προϊόντων σε μορφή αναγνώσιμη από τις μηχανές [36]. Μια οντολογία προϊόντων περιγράφει τις ιδιότητες του προϊόντος όπως την εμφάνιση, τη δομή, τη συμπεριφορά και τη λειτουργικότητα.

Το σύνολο των ονομάτων χαρακτηριστικών που περιγράφουν ένα προϊόν μπορεί να ονομαστεί *σχήμα προϊόντος (product schema)*, ακολουθώντας την ορολογία των σχεσιακών βάσεων. Μπορεί περαιτέρω να γίνει διαφοροποίηση μεταξύ *τοπικού σχήματος (local schema)* (που είναι ένα σύνολο ιδιοτήτων συγκεκριμένου προμηθευτή) και ενός *ολικού σχήματος προϊόντος (global schema)*(που είναι ένα τυποποιημένο και κοινά χρησιμοποιούμενο σχήμα). Η λίστα των ιδιοτήτων στο πρότυπο eCl@ss (που περιγράφεται στην Ενότητα 4.1.1) είναι ένα παράδειγμα ολικού σχήματος προϊόντος.

Βασικά, υπάρχουν δυο διαφορετικοί τρόποι για τον χειρισμό ετερογενών περιγραφών προϊόντων [36] : (1) η δημιουργία ενός προτύπου ή (2) η οικοδόμηση ενός επιπέδου που αντιστοιχίζει διαφορετικά πρότυπα.

Ένα βασικό πρόβλημα με τις παραδοσιακές προσεγγίσεις ταξινόμησης (δηλαδή, eCl@ss [14], UNSPSC [13]) είναι ότι στο περιβάλλον του Σημασιολογικού Ιστού το ίδιο έγγραφο πρέπει να είναι αναγνώσιμο από ένα μεγάλο αριθμό διαφορετικών εταίρων για μια πληθώρα σκοπών. Με άλλα λόγια, ο αποδέκτης και ο χρήστης των δεδομένων δεν είναι προκαθορισμένα, πράγμα που καθιστά δύσκολο να επιτευχθεί συναίνεση για τις κατάλληλες κλάσεις προϊόντων που

αναπαριστούν τη φύση και τις δυνατές χρήσεις ενός προϊόντος, με επαρκείς λεπτομέρειες.

Οι οντολογίες δημιουργήθηκαν για να υποστηρίξουν απαιτήσεις *αναπαράστασης (represenatation)* προϊόντων, σε αντίθεση με τις απαιτήσεις *παρουσίασης (presentation)* των εφαρμογών που περιλαμβάνουν κάποια μορφή ταξινόμησης των προϊόντων και των υπηρεσιών. Οι αναπαραστάσεις των οντολογιών είναι σημασιολογικά σταθερές, συνεπείς (αν και προφανώς πάντα ατελείς γιατί πρόσθετες βελτιώσεις μπορούν πάντα να γίνουν), ελεγχόμενες, τμηματοποιημένες (modular), επαναχρησιμοποιήσιμες και παρέχουν κάποια υποστήριξη των αναγκών παρουσίασης των εφαρμογών.

Η αποτελεσματική επικοινωνία μεταξύ μηχανών είναι μια απαίτηση για το B2B ηλεκτρονικό εμπόριο και ο Σημασιολογικός Ιστός υπόσχεται να κάνει την μεγάλη ποσότητα των δεδομένων στο Διαδίκτυο αναγνώσιμο και επεξεργάσιμο από μηχανές με την τυποποίηση της σημασιολογίας. Δεδομένου ότι τα προϊόντα και οι υπηρεσίες είναι τα βασικά αντικείμενα του εμπορίου, η αναπαράστασή τους σε μορφή αναγνώσιμη από τις μηχανές αποτελεί βασική πρόκληση στο δρόμο για τις επιχειρηματικές εφαρμογές για τον Σημασιολογικό Ιστό.

Η πιο προφανής διαδικασία που θα επωφελούνταν από μια αναγνώσιμη από τις μηχανές αναπαράσταση προϊόντων και υπηρεσιών, όπως είναι οι οντολογίες, είναι η αναζήτηση για πιθανές αντιστοιχίσεις μεταξύ αγοραστών και πωλητών και η εξεύρεση της καλύτερης (ακόμα και αν δεν υπάρχει το τέλειο ταίριασμα) το οποίο είναι γνωστό ως *ταίριασμα (matchmaking)*. Όμως, υπάρχουν πολύ περισσότερες διεργασίες που θα επωφελούνταν από μια αναπαράσταση προϊόντων μέσω οντολογιών στον Σημασιολογικό Ιστό. Τέτοια παραδείγματα είναι [36]:

1. τα *συστήματα συστάσεων (recomender systems)*, τα οποία χρειάζονται σημεία αναφοράς για συγκεκριμένες κατηγορίες προϊόντων ή μοντέλα προϊόντων για να αποθηκεύσουν την γνώση του πεδίου εφαρμογής τους

2. η έκφραση, η τυποποίηση, η μεταφορά ακόμη και η εμπορία γνώσεων για τη σχέση των προϊόντων (π.χ. “τα καρβουνάκια πωλούνται καλά με τα λουκάνικα σχάρας”)
3. οι αναλυτικές εργασίες όπως η λογιστική κόστους, η ανάλυση εξόδων και η συγκριτική αξιολόγηση (benchmarking)
4. η έκφραση μιας απαίτησης όταν γίνονται επερωτήσεις σε μια μηχανή αναζήτησης (π.χ. “Χρειάζομαι ένα μέρος τοποθέτησης που δεν είναι διαβρωτικό και μπορεί να χωρέσει 500 kg.”) και
5. η ενοποίηση δεδομένων καταλόγου που προέρχονται από διαφορετικούς προμηθευτές από κατανεμημένες πηγές (π.χ. απόκτηση πληροφορίας που λείπει με εργασίες συμπερασμού από έγγραφα του Σημασιολογικού Ιστού).

Το B2B ηλεκτρονικό εμπόριο χρειάζεται τις οντολογίες [37]. Κατ’αρχήν υπάρχει η *πληροφοριακή* ανάγκη : επειδή η οντολογία είναι ένα δομημένο εννοιολογικό μοντέλο του πεδίου εφαρμογής του ηλεκτρονικού εμπορίου, υποστηρίζει την βασισμένη σε παράμετρο/ιδιότητα αναζήτηση και πλοήγηση, χρησιμοποιώντας γνώση για προϊόντα και υπηρεσίες από υποψήφιους αγοραστές, για να ανακαλύψουν τι να αγοράσουν, και στην συνέχεια να καθορίσουν τις τιμές και την διαθεσιμότητα. Σε αυτή την περίπτωση, η σχετικά στατική γνώση των οντολογιών αντιστοιχίζεται στα σχετικά δυναμικά δεδομένα των πωλητών. Επιπλέον, μια οντολογία μπορεί να μοντελοποιήσει όχι μόνο εμπορεύματα, αλλά και πράκτορες, δηλαδή αγοραστές και πωλητές, τόσο ανθρώπινους όσο και τεχνητούς. Με την χρησιμοποίηση της γνώσης αναφορικά με τον ρόλο του χρήστη (που μερικές φορές ονομάζεται προφίλ χρηστών ή εξατομίκευση) βελτιώνεται η διαδικασία αναζήτησης, οι επερωτήσεις μπορούν να προσαρμοστούν στις λειτουργίες και ενδιαφέροντα του χρήστη, από πληροφορίες που βασίζονται σε προηγούμενη αλληλεπίδραση με αυτόν.

Το ηλεκτρονικό εμπόριο χρειάζεται επιπλέον τις οντολογίες για *συναλλακτικούς (transactional)* σκοπούς : η γνώση για την οργανωτική δομή μιας επιχείρησης, τη ροή εργασιών, τις διαδικασίες και τα προϊόντα/υπηρεσίες μπορούν να χρησιμοποιηθούν για να βοηθήσουν στην άμεση αγορά και πώληση.

Η χρήση των οντολογιών προϊόντων μπορεί να βελτιώσει την αναζήτηση των προϊόντων στον Ιστό, αλλά και τον χειρισμό των δεδομένων προϊόντων σε B2B σενάρια ηλεκτρονικού εμπορίου.

Τελευταία, οι τεχνολογίες του Σημασιολογικού Ιστού έχουν ωριμάσει ώστε να κάνουν τις αλληλεπιδράσεις στο ηλεκτρονικό εμπόριο πιο ευέλικτες και αυτοματοποιημένες. Ο Σημασιολογικός Ιστός παρέχει σαφή σημασιολογική περιγραφή στις πληροφορίες που είναι διαθέσιμες στο διαδίκτυο, έτσι ώστε να έχουμε αυτοματοποιημένη επεξεργασία και ολοκλήρωση πληροφοριών με βάση την υποκείμενη οντολογία. Η οντολογία ορίζει τους όρους που χρησιμοποιούνται για την περιγραφή ενός τομέα γνώσης η οποία μοιράζεται μεταξύ ανθρώπων, βάσεων δεδομένων και εφαρμογών. Ειδικότερα, η οντολογία κωδικοποιεί γνώση που μπορεί να διαπερνά πολλά πεδία εφαρμογής, όπως επίσης και να περιγράψει τις σχέσεις μεταξύ των πεδίων αυτών.

Ο Πίνακας 3.1 συνοψίζει τη συμβολή της οντολογίας σε κάποια τυπικά προβλήματα των εικονικών αγορών [8].

Πίνακας 3.1: Επισκόπηση συνεισφοράς οντολογιών στις εικονικές αγορές

Λειτουργία	Πρόβλημα στην εικονική αγορά	Συμβολή οντολογίας
Ταίριασμα αποφάσεων (matchmaking)	Το ταίριασμα αποφάσεων είναι συχνά αναποτελεσματικό λόγω του "κακού" καθορισμού των προϊόντων με περιορισμένες ιδιότητες.	Μια διαμοιρασμένη και συμφωνημένη οντολογία προσφέρει κοινό, ευέλικτο και επεκτάσιμο ορισμό των προϊόντων και των απαιτήσεων για ταίριασμα αποφάσεων (matchmaking) και των μετέπειτα επιχειρησιακών διαδικασιών.
	Είναι δύσκολο να καθοριστούν οι απαιτήσεις για ένα σύνθετο προϊόν γιατί οι συσχετίσεις μεταξύ των χαρακτηριστικών και των τιμών τους αγνοούνται.	Περίπλοκες απαιτήσεις μπορούν να αναλυθούν σε απλές έννοιες για τον εξορθολογισμό της εκμαίευση των επιλογών.

	Οι αλληλεπιδράσεις των χρηστών περιορίζονται σε χειροκίνητες κυρίως, το οποίο είναι χρονοβόρο.	Η πρόσβαση γίνεται από αυτοματοποιημένους πράκτορες μέσω προδιαγραφών Σημασιολογικού Ιστού για περισσότερες επιχειρηματικές ευκαιρίες.
Συστάσεις	Οι συστάσεις είναι συνήθως εφικτές μόνο μέσα στην ίδια κατηγορία.	Η Οντολογία βοηθά στην εξαγωγή εναλλακτικών συστάσεων.
	Για την αξιολόγηση είναι απαραίτητη η προρυθμισμένη σύνθεση για κάθε τύπο προϊόντος.	Η οντολογία βοηθά στις συστάσεις με την αξιολόγηση των προσφορών όσον αφορά την ευέλικτη συνολική κλιμάκωση.
	Η διαγώνια πώληση (cross-sale) και η ομαδοποίηση των αγοραστών και των πωλητών που έχουν παρόμοια αιτήματα είναι δύσκολη.	Ασορτί ομαδοποίηση των αγοραστών και των πωλητών όπως και η διαγώνια πώληση είναι δυνατή κατά τεκμήριο με την οντολογία.
Διαπραγμάτευση	Δεν μπορεί να γίνει υπονοούμενη παραγγελία εναλλακτικών λύσεων.	Υπονοούμενη παραγγελία εναλλακτικών λύσεων που εξάγεται μέσω κληρονομικότητας.
	Χειροκίνητη διαπραγμάτευση ή ανεπαρκής υποστήριξη διαπραγμάτευσης προκαλούν αναποτελεσματική διαδικασία και αναποτελεσματική αναγνώριση.	Η σημασιολογία που είναι κατανοητή από την μηχανή διευκολύνει την διαπραγμάτευση και τον αυτόματο προσδιορισμό των προϊόντων και των υπηρεσιών.

3.3 Μάθηση Οντολογίας από Σχεσιακά Σχήματα

Οι data-intensive ιστοσελίδες (σε αντίθεση με τις στατικές ιστοσελίδες) είναι εκείνες που αλλάζουν δυναμικά κατά τη διάρκεια των αναζητήσεων/επερωτήσεων χρήστη και υλοποιούνται τυπικά μέσω σχεσιακών βάσεων. Μια από τις πιο συχνές εφαρμογές για data-intensive ιστοσελίδες είναι

οι εφαρμογές Ηλεκτρονικού Εμπορίου. Οι σελίδες αυτές χαρακτηρίζονται από αυτόματη ενημέρωση των περιεχομένων τους και απλοποιημένη συντήρηση του σχεδιασμού τους. Όμως παρουσιάζουν δυο περιορισμούς. Πρώτο, δημιουργούν έναν κρυφό Ιστό γιατί το περιεχόμενό τους δεν είναι εύκολα προσβάσιμο από κανένα αυτόματο εργαλείο επεξεργασίας, συμπεριλαμβανομένων των ρομπότ ευρετηρίασης των μηχανών αναζήτησης (search engine indexing robots). Δεύτερο, το περιεχόμενο των προσανατολισμένων σε βάσεις δεδομένων (database-driven) ιστοσελίδων που περιγράφονται με χρήση της HTML δεν είναι αναγνώσιμο από τις μηχανές. Οι οντολογίες υπόσχονται να λύσουν αυτό το δεύτερο πρόβλημα, όμως υπάρχει το πρόβλημα του πώς θα κατασκευαστούν οντολογίες γιατί η κατασκευή τους είναι δαπανηρή και αυτό αποτελεί εμπόδιο στην πρόοδο της δραστηριότητας του Σημασιολογικού Ιστού. Η χειρωνακτική κατασκευή των οντολογιών είναι δύσκολη, χρονοβόρα και επιρρεπής σε λάθη και κυρίως προκαλεί συμφόρηση στην απόκτηση γνώση. Τα πλήρως αυτοματοποιημένα εργαλεία είναι ακόμη σε πολύ αρχικό στάδιο για να υλοποιηθούν. Γι αυτό, η ημιαυτόματη εξαγωγή οντολογιών είναι μια πρακτική βραχυπρόθεσμη λύση, που επιτρέπει την εξαγωγή γνώσης από data-intensive εφαρμογές Ιστού. Για την υλοποίηση αυτής της διαδικασίας συνήθως χρησιμοποιούνται τεχνικές reverse engineering, που ορίζεται ως μια διαδικασία ανάλυσης ενός πρότυπου (legacy) συστήματος για τον εντοπισμό όλων των μερών του και των σχέσεων μεταξύ τους.

Ωστόσο, επειδή οι οντολογίες αποτελούν ένα σχετικά νέο πεδίο έρευνας, υπάρχουν λίγες προσεγγίσεις που θεωρούν τις οντολογίες ως στόχο για reverse engineering τεχνικές. Αυτές οι προσεγγίσεις συνήθως απαιτούν ως είσοδο περισσότερη πληροφορία από ότι είναι δυνατό να δοθεί στην πράξη, καθώς η ολοκληρωμένη πληροφορία για μια σχεσιακή βάση δεδομένων συνήθως δεν είναι δημόσια διαθέσιμη. Επίσης, στις προσεγγίσεις αυτές γίνονται μη ρεαλιστικές υποθέσεις για τα δεδομένα εισόδου, για παράδειγμα ότι η σχεσιακή βάση είναι σε 3NF μορφή. Σε μια προσπάθεια να αποφευχθούν αυτοί οι περιορισμοί προτείνονται άλλες προσεγγίσεις για reverse engineering data-intensive εφαρμογών Ιστού στον Σημασιολογικό Ιστό που βασίζονται στις HTML σελίδες.

Όπως είπαμε και νωρίτερα, ενώ έχουν πραγματοποιηθεί πολλές έρευνες για reverse engineering σχεσιακών σχημάτων που προτείνουν μεθόδους και κανόνες για εξαγωγή μοντέλου οντοτήτων συσχετίσεων (entity relationship model) και αντικειμενοστρεφών μοντέλων (object models) από σχεσιακές βάσεις, είναι λιγότερες οι έρευνες που θεωρούν τις οντολογίες ως τον στόχο για reverse engineering. Αυτές οι προσεγγίσεις χωρίζονται κατά βάση σε πέντε κατηγορίες:

1. *Προσεγγίσεις που βασίζονται στην ανάλυση των επερωτήσεων χρηστών (analysis of user queries)* : Η προσέγγιση στο [38] δημιουργεί μια οντολογία από το σχεσιακό σχήμα, η οποία στη συνέχεια βελτιστοποιείται από τις επερωτήσεις του χρήστη. Αυτή η προσέγγιση δεν δημιουργεί αξιώματα (axioms), που είναι μέρος των οντολογιών. Επίσης, δεν αναλύονται πάντα τα σημασιολογικά χαρακτηριστικά του σχεσιακού σχήματος.
2. *Προσεγγίσεις που βασίζονται στην ανάλυση του σχεσιακού σχήματος (analysis of relational schema)* : Η προσέγγιση στο [39] παρέχει ένα σύνολο κανόνων για την αντιστοίχιση δομών του σχεσιακού σχήματος σε σημασιολογικά ισοδύναμες δομές στην οντολογία. Αυτοί οι κανόνες βασίζονται σε ανάλυση των σχέσεων, κλειδιών και εξαρτήσεις συμπερίληψης (inclusion dependencies). Στο [40] προτείνεται η αυτοματοποίηση της διαδικασίας της συμπλήρωσης στιγμιοτύπων και των τιμών των ιδιοτήτων τους σε μια οντολογία, χρησιμοποιώντας τα δεδομένα που εξάγωνται από εξωτερικές σχεσιακές πηγές. Αυτή η μέθοδος χρησιμοποιεί μια διεπαφή μεταξύ της οντολογίας και της πηγής δεδομένων, μοντελοποιημένη στην οντολογία και υλοποιημένη σε XML σχήμα. Απαιτεί πολλές συνιστώσες : οντολογία, XML σχήμα, μεταφραστής XML.
3. *Προσεγγίσεις που βασίζονται στην ανάλυση των πλειάδων (analysis of tuples)* : Η προσέγγιση στο [41] δημιουργεί μια οντολογία βασισμένη στην ανάλυση του σχεσιακού σχήματος. Επεδή το σχεσιακό σχήμα συχνά έχει λίγη ρητή σημασιολογία, αυτή η προσέγγιση αναλύει επίσης τις πλειάδες σε μια σχεσιακή βάση για την ανακάλυψη πρόσθετης “κρυμμένης” σημασιολογίας

(π.χ. κληρονομικότητα). Ωστόσο, αυτή η προσέγγιση είναι πολύ χρονοβόρα συγκριτικά με τον αριθμό των πλειάδων στην σχεσιακή βάση δεδομένων.

4. *Προσεγγίσεις που βασίζονται στην ανάλυση των HTML-πινάκων (analysis of HTML-table)* : Η προσέγγιση στο [42] βασιζόμενη στην τεχνική εξαγωγής εννοιολογικής μοντελοποίησης προσπαθεί να καταλάβει την δομή ενός πίνακα και το εννοιολογικό περιεχόμενο, να ανακαλύψει τους περιορισμούς που υπάρχουν μεταξύ εννοιών που εξάγονται από τον πίνακα, να αντιστοιχίσει τις εντοπισμένες έννοιες με αυτές από μια πιο γενική περιγραφή των σχετικών εννοιών, και να συγχωνεύσει το αποτέλεσμα της δομής με άλλες παρόμοιες αναπαραστάσεις γνώσης. Αυτή η προσέγγιση απαιτεί βοηθητικές πληροφορίες που περιλαμβάνει λεξικά και λεξιλογικά δεδομένα (δηλαδή WordNet, αναλυτή Φυσικής Γλώσσας, και βιβλιοθήκες πλαισίων δεδομένων).

5. *Προσεγγίσεις που βασίζονται στην ανάλυση HTML-φορμών (analysis of HTML-forms) :* Η προσέγγιση στο [43] κατασκευάζει μια οντολογία βασισμένη στην ανάλυση των HTML – φορμών για την εξαγωγή form model schema, την μετατροπή του σε οντολογία και την δημιουργία στιγμιοτύπων οντολογίας από δεδομένα που υπάρχουν στις σελίδες. Το βασικό μειωνέκτημα αυτής της προσέγγισης είναι ότι δεν παρέχει τρόπο για τον εντοπισμό σχέσεων κληρονομικότητας.

Για την αποφυγή των προβλημάτων που έχουν οι παραπάνω προσεγγίσεις το [44] επεκτείνει την προσέγγιση που βασίζονται σε ανάλυση HTML-φορμών και βασίζεται στην ιδέα ότι η σημασιολογική πληροφορία μιας σχεσιακής βάσης μπορεί να εξαχθεί με την ανάλυση των σχετικών HTML σελίδων. Αυτή η σημασιολογία εμπλουτίζεται με αυτή που εξάγεται από το σχεσιακό σχήμα ώστε να κατασκευαστεί η οντολογία. Η χρήση HTML-φορμών ως πηγή για την δημιουργία οντολογιών έχει πολλά θετικά. Οι HTML φόρμες είναι κατάλληλες διεπαφές για εισαγωγή, μετατροπή και θεώρηση των δεδομένων των ιστοσελίδων. Έτσι η ανάλυσή τους μπορεί να δώσει σημαντικές πληροφορίες όπως είναι τα υποχρεωτικά (mandatory) δεδομένα και οι δυνατές τιμές των

δεδομένων. Τα δεδομένα σε μια φόρμα είναι συνήθως δομημένα, ενώ η δομή μιας σχεσιακής βάσης δεν είναι γνωστή και συνήθως δεν είναι δημόσια διαθέσιμη. Επίσης, οι HTML φόρμες αναπαριστούν μερικώς την λογική δομή μιας σχεσιακής βάσης που μας ενδιαφέρει, παρά τη φυσική δομή της (δηλαδή το σχεσιακό σχήμα). Παρέχει μια φιλική προς τον χρήστη διεπαφή με την σχεσιακή βάση, αποκρύπτοντας ότι το σχήμα της μπορεί να μην είναι καλώς σχεδιασμένο, ή βελτιστοποιημένο ή ακόμη και κανονικοποιημένο. Επίσης τα ονόματα των πεδίων στις HTML φόρμες είναι πιο σαφή με περισσότερο νόημα από τα αντίστοιχα ονόματα σχέσεων και ιδιοτήτων στο σχεσιακό σχήμα. Τέλος, οι HTML φόρμες συνήθως περιλαμβάνουν οδηγίες και πρόσθετες πληροφορίες στο πώς διαχειρίζονται τα δεδομένα και πώς είναι δομημένα.

ΚΕΦΑΛΑΙΟ 4

ΠΑΡΟΥΣΙΑΣΗ ΟΝΤΟΛΟΓΙΩΝ ΠΡΟΪΟΝΤΩΝ

Ένα σημαντικό μέρος των δεδομένων και της διαχείρισης περιεχομένου στα σενάρια του ηλεκτρονικού επιχειρείν έχει να κάνει με την ανταλλαγή δεδομένων σχετικά με προϊόντα, μεταξύ επιχειρηματικών οντοτήτων και την ενσωμάτωσή τους σε εφαρμογές-στόχους (π.χ. ERP συστήματα) ή έγγραφα-στόχους (π.χ. ηλεκτρονικούς καταλόγους) στην πλευρά του παραλήπτη [9]. Εργασίες ολοκλήρωσης (integration) περιεχομένου μπορούν να αυτοματοποιηθούν καλύτερα εάν οι περιγραφές κειμένου επαυξάνονται με μια σημασιολογική πληροφορία που να είναι αναγνώσιμη από τις μηχανές. Για το σκοπό αυτό, τα πρότυπα κατηγοριοποίησης προϊόντων και υπηρεσιών όπως το UNSPSC [13], eCl@ss [14], eOTD [27], ή το Rosettanet Technical Dictionary (RNTD) [26] χρησιμοποιούνται ευρέως.

4.1 Πρότυπα Κατηγοριοποίησης

Υπάρχουν αμέτρητες προσεγγίσεις για την κατηγοριοποίηση των αγαθών, που κυμαίνονται από αρκετά χονδροειδής ταξινομίες, δημιουργημένες για ειδικές περιπτώσεις και στατιστικές των οικονομικών δραστηριοτήτων, όπως είναι το North American Industry Classification System (NAICS) και ο προκάτοχός του SIC, μέχρι εκφραστικές γλώσσες περιγραφής προϊόντων και υπηρεσιών, όπως είναι eCl@ss, eOTD, ή RNTD. Το UNSPSC αναφέρεται ευρέως ως ένα παράδειγμα οντολογίας προϊόντος και είναι μεταξύ αυτών των δυο άκρων, παρέχοντας μια ανεξαρτήτως-βιομηχανίας ταξινομία κατηγοριών προϊόντων και υπηρεσιών, αλλά δεν έχει τυποποιημένες ιδιότητες για την λεπτομερή περιγραφή των προϊόντων. Μπορεί κανείς να πει ότι το UNSPSC, το eCl@ss και το eOTD είναι σήμερα τα πιο σημαντικά οριζόντια πρότυπα (δηλαδή, που καλύπτουν ένα ευρύ φάσμα βιομηχανιών) και το RNTD έχει υψηλό βαθμό λεπτομέρειας, αν και περιορίζεται σε ένα μικρό τμήμα των προϊόντων (ηλεκτρονικά και IT εξαρτήματα).

Όλα τα πρότυπα αποτελούνται από κάποια στοιχεία [9]:

- *Κλάσεις προϊόντων* : Η ομαδοποίηση των προϊόντων σε κατηγορίες επηρεάζεται από τον σκοπό του προτύπου. Για παράδειγμα, οι κατηγορίες συλλέγουν προϊόντα με βάση τη *φύση* των προϊόντων ή με βάση τη *χρήση* τους.
- *Ιεραρχία των κλάσεων* : Η ιεραρχία των κλάσεων είναι άμεσα συνδεδεμένη με τη χρήση του προτύπου. Για παράδειγμα το eCl@ss σχεδιάστηκε με την ιδέα της ομαδοποίησης προϊόντων από τη σκοπιά του *οργανισμού-αγοραστή* ή του *διευθυντή προμηθειών*.
- *Λεξικό ιδιοτήτων* : Περιγράφει αναλυτικά τις ιδιότητες της οντολογίας, με το πεδίο τιμών, τις μονάδες μέτρησης και με αναφορές σε διεθνή πρότυπα.
- *Απαρίθμηση τιμών ιδιοτήτων* : Κάποιες ιδιότητες παίρνουν τιμές από ένα συγκεκριμένο σύνολο τιμών το οποίο αναφέρεται σε αυτή την ξεχωριστή συλλογή.
- *Συσχέτιση κλάσεων- ιδιοτήτων* : Αναφέρονται οι ιδιότητες που σχετίζονται με κάθε κλάση. Ανάλογα το πρότυπο αυτή η περιγραφή μπορεί να είναι πλήρης ή πιο χαλαρή.
- *Λέξεις κλειδιά* : Σύνολα λέξεων-κλειδιών και των σχέσεων μεταξύ τέτοιων λέξεων και κατηγοριών ή ιδιοτήτων. Είναι χρήσιμες για χειροκίνητες αναζητήσεις για σωστές καταχωρήσεις.

Λόγω της συνεχούς καινοτομίας στο πεδίο των προϊόντων και των υπηρεσιών, όλα τα πρότυπα είναι έργα σε εξέλιξη με πολλές εκδόσεις κάθε χρόνο.

Στις επόμενες δυο ενότητες θα γίνει μια περιγραφή των δυο πιο δημοφιλών προτύπων κατηγοριοποίησης, eCl@ss και UNSPSC. Επίσης, θα περιγραφούν τα αποτελέσματα των δυο προτύπων σε σχέση με τις παρακάτω μετρικές [9] :

(1) *Μέγεθος και ανάπτυξη* : Η μετρική αυτή αντικατοπτρίζει το μέγεθος του λεξιλογίου, δηλαδή τον αριθμό των γενικών εννοιών προϊόντων και υπηρεσιών και πώς αυτό αλλάζει με την πάροδο του χρόνου. Επίσης δείχνει το βαθμό της δυναμικής αλλαγής μεταξύ δυο οποιονδήποτε διαδοχικών εκδόσεων, το οποίο

είναι σημαντικό για σταθερούς χρήστες, δεδομένου ότι βοηθά στον καθορισμό μιας κατάλληλης στρατηγικής για να αντιμετωπίσουν τις αλλαγές στις εκδόσεις. Οι τροποποιημένες κλάσεις συχνά απαιτούν χειροκίνητο έλεγχο για το κατά πόσο οι υπάρχουσες αναθέσεις κλάσεων εξακολουθούν να ισχύουν.

Για μια καλή κάλυψη των εννοιών που απαιτούνται σε κάποιο τομέα, οποιοδήποτε πρότυπο απαιτεί έγκαιρη και πλήρη ανάδραση από την κοινότητα των χρηστών για καταχωρήσεις που λείπουν και μια βελτιστοποιημένη διαδικασία τυποποίησης που κάνει εγκαίρως τα αντίστοιχα νέα στοιχεία διαθέσιμα.

(2) *Ιεραρχία και ισορροπία περιεχομένου* : Αυτές οι μετρικές δείχνουν πώς η κατανομή των κλάσεων αναπτύχθηκε με την πάροδο του χρόνου κατά μήκος των τμημάτων, ώστε να φανεί εάν το δεδομένο πρότυπο γίνεται όλο και πιο ισορροπημένο ή εάν ο βαθμός ανισορροπίας αυξάνεται. Επίσης, δεδομένου ότι ο συντελεστής μεταβλητότητας μπορεί να χρησιμοποιηθεί για να συγκρίνουμε κατανομές με διαφορετική μέση τιμή, είναι ένας καλός δείκτης για τη σύγκριση πολλών προτύπων.

Για οριζόντια πρότυπα προϊόντων και υπηρεσιών, αυτό αποκαλύπτει αν το πρότυπο είναι μια πραγματικά οριζόντια προσέγγυση ή είναι οριζόντια μόνο σε σχέση με την ύπαρξη των τμημάτων, αλλά σε πιο αναλυτικό επίπεδο μάλλον είναι επικεντρωμένο αρκετά κάθετα. Ένα αληθινά οριζόντιο πρότυπο δεν απαιτεί μόνο την ύπαρξη των τμημάτων για ένα ευρύ φάσμα εννοιών αλλά και πραγματικές καταχωρήσεις στις βαθύτερες διακλαδώσεις όλων των τμημάτων.

(3) *Βιβλιοθήκη ιδιοτήτων* : Το μέγεθος της βιβλιοθήκης ιδιοτήτων αντικατοπτρίζει το πλήθος των εννοιών που αφορούν ιδιότητες για ένα δεδομένο πρότυπο. Ωστόσο, μπορεί να υποπτευθεί ότι ο πλεονασμός είναι μεγάλο πρόβλημα όσον αφορά τις ιδιότητες, γιατί η συχνά κατανεμημένη ανάπτυξη των προτύπων κάνει πολύ πιθανή την δημιουργία πλεονοζόντων ιδιοτήτων, όταν η ύπαρξη μιας ισοδύναμης ιδιότητας δεν έχει συνειδητοποιηθεί λόγω διαφορετικών συμβάσεων ορολογίας. Στο παρόν στάδιο, αυτό αποτελεί μάλλον μια χονδροειδή μετρική, καθώς δεν δείχνει το ποσό της εργασίας ενοποίησης (π.χ. την διαγραφή των

πλεονάζοντων ιδιοτήτων). Εάν ο αριθμός των ιδιοτήτων έχει μειωθεί, ενώ ο αριθμός των κλάσεων έχει αυξηθεί, μπορούμε να υποθέσουμε ότι έχει πραγματοποιηθεί κάποια εξυγείανση.

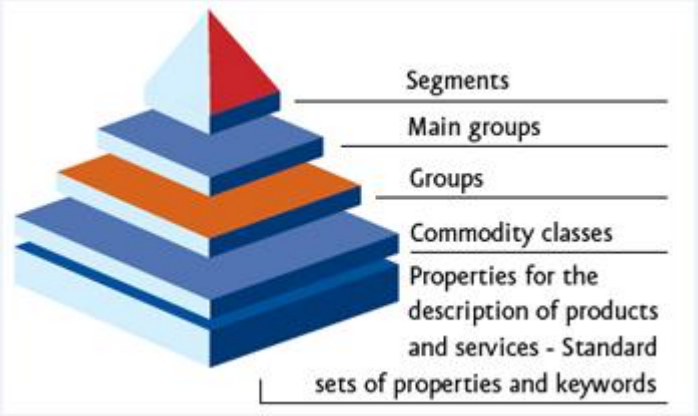
Είναι ιδιαίτερα επιθυμητό να έχουν οριστεί κατάλληλοι λεξιλογικοί χώροι για όλες τις ιδιότητες, δηλαδή απαριθμημένοι τύποι δεδομένων για ιδιότητες που δεν μπορούν να αναπαρασταθούν με σαφήνεια χρησιμοποιώντας τους στάνταρ τύπους δεδομένων. Ωστόσο, παρατηρούμε συχνά ότι τέτοιοι ορισμοί ιδιοτήτων είναι ατελείς (π.χ. ιδιότητα που ορίζεται ως μια αλφαρηθμητική ακολουθία με λιγότερο από 30 χαρακτήρες). Αυτό εμποδίζει την αυτόματη ερμηνεία των τιμών των ιδιοτήτων.

(4) *Ποιότητα των συνόλων ιδιοτήτων συγκεκριμένων κλάσεων* : Οι λίστες ιδιοτήτων λένε σε έναν χρήστη ενός προτύπου ποιες ιδιότητες πρέπει να χρησιμοποιηθούν για να περιγράψει ένα μοντέλο προϊόντος με λεπτομέρεια. Αυτές οι συστάσεις είναι μέρος πολλών προτύπων και πρέπει να περιέχουν όλες τις απαραίτητες ιδιότητες, αλλά όχι μια άγρια συλλογή από κάθε χρησιμοποιούμενη ιδιότητα, γιατί αυτό κάνει την αυτόματη επεξεργασία των συγκεκριμένων προϊόντων δύσκολο, καθώς στοιχεία του ίδιου τύπου θα μπορούσαν να περιγραφούν χρησιμοποιώντας διαφορετικές ιδιότητες. Η δημιουργία και η διατήρηση τέτοιων συνόλων ιδιοτήτων για κάθε κατηγορία είναι ένα τεράστιο έργο, γιατί απαιτεί συναίνεση σε ένα πολύ αναλυτικό επίπεδο. Η παροχή ιδιοτήτων συχνά θεωρείται ως στοιχείο διάκρισης μεταξύ των προτύπων, αλλά μέχρι στιγμής έχει χρησιμοποιηθεί σε ένα διαρθρωτικό επίπεδο, δηλαδή εάν το μοντέλο δεδομένων του προτύπου υποστηρίζει ιδιότητες, και όχι εάν το πρότυπο πραγματικά περιέχει συγκεκριμένες αναθέσεις ιδιοτήτων. Οι μετρικές σε αυτή την ενότητα αποκαλύπτουν το βαθμό στον οποίο διάφορα πρότυπα πραγματικά υλοποιούν σύνολα ιδιοτήτων. Επίσης, μόνο το πλήθος των συγκεκριμένων αναθέσεων ιδιοτήτων καθορίζει το ποσό της προόδου στην δημιουργία ολοκληρωμένων εννοιών προϊόντων και υπηρεσιών.

4.2.1 Το πρότυπο κατηγοριοποίησης eCl@ss

Το eCl@ss είναι ένα διεπιχειρησιακό πρότυπο που αναπτύχθηκε στην Γερμανία για την ταξινόμηση και σαφή περιγραφή προϊόντων και υπηρεσιών. Έχει καθιερωθεί ως το μοναδικό πρότυπο συμβατό με το ISO/IEC βιομηχανικό πρότυπο σε εθνικό και διεθνές επίπεδο. Από το 2000, το eCl@ss κατέχει μια καθιερωμένη θέση στη βιομηχανία, το εμπόριο, τις τέχνες, τα τρόφιμα, τις υπηρεσίες και πολλά άλλα. Με τις 38.000 κλάσεις προϊόντων και τις 16.000 ιδιότητες, το eCl@ss καλύπτει την πλειοψηφία των εμπορεύσιμων αγαθών και υπηρεσιών. Πολλά βιομηχανικά πρότυπα (π.χ. από την ηλεκτρονική βιομηχανία, την ιατρική τεχνολογία, τις κατασκευές, τη βιομηχανία χαρτιού) επιζητούν την διαλειτουργικότητα για να αξιοποιηθούν οι δυνατότητες ενός διεπιχειρησιακού προτύπου [14].

Η ομαδοποίηση υλικών, προϊόντων και υπηρεσιών γίνεται σύμφωνα με μια λογική δομή σε μια ιεραρχία με ένα επίπεδο λεπτομέρειας που αντιστοιχεί στις ιδιότητες του συγκεκριμένου προϊόντος, που μπορούν να περιγραφούν χρησιμοποιώντας ιδιότητες που υπακούουν σε κάποιους κανόνες. Τα προϊόντα και οι υπηρεσίες μπορούν να κατανεμηθούν σε τέσσερα επίπεδα όπως φαίνεται στο Σχήμα 4.1.



Σχήμα 4.1 : Το τεσσάρων επιπέδων ιεραρχικό σύστημα ταξινόμησης eCl@ss.

Κάθε επίπεδο προσθέτει ένα 2-συμβόλων πρόθεμα στον eCl@ss κωδικό, που όλο μαζί σχηματίζει ένα 8-ψήφιο κωδικό. Εκτός από την ταξινόμηση το eCl@ss

παρέχει για κάθε τάξη στην ιεραρχία της κατάταξης, τη λεγόμενη τάξη εφαρμογής, η οποία χαρακτηρίζεται από ορισμένες συγκεκριμένες ιδιότητες, που μπορούν να χρησιμοποιηθούν για την περιγραφή του ταξινομηθέντος στοιχείου. Τα περιεχόμενα του eCl@ss είναι διαθέσιμα σε πολλές γλωσσές (π.χ. Κινέζικα, Γαλλικά, Ιταλικά, Γιαπωνέζικα, Ισπανικά, Τουρκικά, Ρωσικά, Πορτογαλλικά κ.ά.), με τα Αγγλικά και τα Γερμανικά να είναι οι πιο πλήρεις εκδόσεις.

Το πρότυπο eCl@ss βασίζεται σε τμήματα της αγοράς προϊόντων και παρέχει πολλές καταχωρήσεις για τις *χρήσεις* του προϊόντος [15]. Μια σημαντική οπτική κατά την περιγραφή ενός προϊόντος είναι η συνειστώμενη ή δυνατή χρήση του. Για πράδειγμα, το αλάτι (NaCl) μπορεί να χρησιμοποιηθεί στο νοικοκυριό, ως αφυγραντικό, ή για να λιώσει το χιόνι και τους πάγους στους δρόμους. Οι ιδιότητες του προϊόντος μπορεί να είναι ακριβώς οι ίδιες, αλλά αυτές είναι τρείς διαφορετικές χρήσεις. Το δύσκολο ζήτημα είναι ότι αν δύο προϊόντα είναι υποκατάστατα ή όχι συχνά εξαρτάται από την επιδιωκώμενη χρήση τους. Για παράδειγμα, ως αφυγραντικό μπορεί κανείς να χρησιμοποιήσει χλωριούχο κάλιο (KaCl), ενώ ως επιτραπέζιο αλάτι δεν θα μπορούσε. Δηλώσεις σχετικά με τη χρήση ενός προϊόντος είναι πιο χρήσιμες για έναν αγοραστή, επειδή ένας αγοραστής θέλει ένα προϊόν για μια συγκεκριμένη χρήση, όχι βάσει μιας ιδιότητας που έχει. Δηλώσεις σχετικά με την πιθανή χρήση ενός προϊόντος δεν είναι όμως και πολύ χρήσιμες για τους κατασκευαστές και τους πωλητές, γιατί μπορεί να μην γνωρίζουν τι κάνουν οι πελάτες τους με ένα προϊόν. Ο ίδιος τροχός, για παράδειγμα, θα μπορούσε να χρησιμοποιηθεί για κάρρα, για παιχνίδια, για φυτά ή μηχανήματα.

Σύμφωνα με τις μετρικές που παρουσιάζονται στο [9] και αναφέρθηκαν στην προηγούμενη ενότητα το πρότυπο eCl@ss παρουσιάζει τα παρακάτω χαρακτηριστικά :

(1) *Μέγεθος και ανάπτυξη* : Το πρότυπο ξεκίνησε στην έκδοση 4.1 με 15315 κλάσεις και ο συνολικός αριθμός των κλάσεων στην έκδοση 5.1de είναι 25658 κλάσεις. Είναι η πιο εύκολα διαθέσιμη μετρική και χρησιμοποιείται συχνά από τους οργανισμούς τυποποίησης για την προώθηση των προτύπων τους. Αυτή η

βασική μετρική ωστόσο δεν αποκαλύπτει το πραγματικό μέγεθος συντήρησης και μπορεί επίσης να είναι πολύ μεροληπτική λόγω του μεγάλου μεγέθους κατηγοριών σε πολύ συγκεκριμένες περιοχές. Πολλά πρότυπα δημιουργήθηκαν με την συγχώνευση υφιστάμενων προτύπων από συγκεκριμένους τομείς (eCl@ss : καλύπτει τις ανάγκες της χημικής βιομηχανίας, eOTD : δημόσιες συμβάσεις του ΝΑΤΟ). Από το πλήθος των νέων και τροποποιημένων κλάσεων για τις εκδόσεις μέχρι την 5.1beta, το eCl@ss έχει κατά μέσο όρο αυξηθεί έως και κατά 280 κλάσεις. Επίσης φαίνεται ότι γίνεται σημαντική συντήρηση των υπαρχόντων καταχωρήσεων.

(2) *Ιεραρχία και ισορροπία περιεχομένου* : Από την σύγκριση του πλήθους των εννοιών που περιέχονται στο μεγαλύτερο και στα τρία μεγαλύτερα τμήματα υψηλότερου επιπέδου, συμπεραίνεται ότι το μεγαλύτερο μερίδιο κατηγοριών προέρχεται από πολύ λίγες διακλαδώσεις και ο βαθμός ανισορροπίας είναι εμφανής.

Το ποσό των κατηγοριών που αναπαριστούν υπηρεσίες στο eCl@ss είναι 1064, δηλαδή ποσοστό 4%. Ο τομέας των υπηρεσιών διαφέρει από την αναπαράσταση απτών προϊόντων, π.χ. επειδή η εκπλήρωσή τους δεσμεύεται από τις ιδιότητες του πελάτη υπηρεσιών, ιδίως σε συνάρτηση με τον τόπο και τον χρόνο. Επίσης, θα μπορούσε να υπάρχουν βιομηχανίες όπου λόγω του μεγάλου όγκου τους, οι υπηρεσίες είναι ειδικού ενδιαφέροντος για ανάλυση εξόδων. Είναι έτσι χρήσιμο να καθορίζεται το ποσοστό των κλάσεων υπηρεσιών.

(3) *Βιβλιοθήκη ιδιοτήτων* : Το πλήθος των ιδιοτήτων στο eCl@ss (5.1de) είναι 5525, εκ των οποίων το 19% περιλαμβάνει απαριθμημένους τύπους δεδομένων.

(4) *Ποιότητα των συνόλων ιδιοτήτων συγκεκριμένων κλάσεων* : Το eCl@ss (5.1de) περιέχει μόνο στο 43% των κλάσεων συγκεκριμένη ανάθεση ιδιοτήτων. Υπάρχουν κλάσεις με μεγάλη διαφορά στο πλήθος των ιδιοτήτων και αυτό δείχνει μόνο μερική πρόοδο στην ανάπτυξη της ανάθεσης ιδιοτήτων.

4.2.2 Το πρότυπο κατηγοριοποίησης UNSPSC

Το πρότυπο United Nations Standard Products and Services Code® (UNSPSC®) παρέχει ένα ανοιχτό, παγκόσμιο, πολλών-τομέων πρότυπο για αποτελεσματική και ακριβή ταξινόμηση προϊόντων και υπηρεσιών [13]. Αποτελείται από μια ιεραρχία πέντε επιπέδων κωδικοποιημένη σαν ένας αριθμός 8-ψηφίων. Το UNSPSC αναπτύχθηκε από κοινού από το Πρόγραμμα Ανάπτυξης του ΟΗΕ και την Dun & Bradstreet Corporation το 1981 και σήμερα την διαχειρίζεται η GS1 ΗΠΑ, που είναι υπεύθυνη για την εποπτεία των αιτήσεων αλλαγής κωδικών, την αναθεώρηση των κωδικών και την έκδοση τακτικά προγραμματισμένων ανανεώσεων κωδικών, καθώς και την διαχείριση ειδικών σχεδίων και πρωτοβουλιών. Η τρέχουσα έκδοση αποτελείται από περισσότερους από 50000 όρους. Το σύνολο των κωδικών είναι διαθέσιμο στις γλώσσες Αγγλικά, Γαλλικά, Γερμανικά, Ισπανικά, Ιταλικά, Γιαπωνέζικα, Κορεάτικα, Ολλανδικά, Κινέζικα, Πορτογαλλικά, Δανέζικα, Νορβηγικά, Σουηδικά και Ουγγαρέζικα.

Το UNSPSC για ένα δεδομένο στοιχείο αποτελείται από πέντε διψήφια αναγνωριστικά, τα οποία μαζί ταξινομούν το στοιχείο στην πέντε-επιπέδων ιεραρχία. Τα πέντε επίπεδα της ταξινόμησης είναι το "Τομέας", η "Οικογένεια", η "Κλάση", το “Εμπόρευμα”, και η “Επιχειρηματική Λειτουργία”. Η “Επιχειρηματική Λειτουργία” είναι προαιρετική.

Παράδειγμα

Τομέας 44. Εξοπλισμός Γραφείου και Υλικά και Αναλώσιμα.

Οικογένεια 10. Μηχανές γραφείου και τα αναλώσιμα και τα αξεσουάρ τους.

Κλάση 15. Φωτοαντιγραφικές μηχανές.

Εμπόρευμα 01. Φωτοτυπικά

Επιχειρηματική Λειτουργία 14. Λιανεμπόριο

Το σχήμα ταξινόμησης προϊόντων στο UNSPSC πρότυπο βασίζεται στην περιγραφή των *χαρακτηριστικών των προϊόντων* [15]. Το χαρακτηριστικό ενός προϊόντος, π.χ. η πρώτη ύλη του ("ανοξείδωτος χάλυβας"), τα χημικά συστατικά του ("NaCl" - αλάτι), ή λεπτομέρειες σχετικά με την διαδικασία παραγωγής ("ιονισμένο") είναι μια από τις πολλές οπτικές που μπορούν να χρησιμοποιηθούν κατά την περιγραφή ενός προϊόντος. Αυτή είναι συνήθως η προτιμώμενη οπτική για τον κατασκευαστή, γιατί τα αντίστοιχα χαρακτηριστικά είναι γνωστά και επίσης είναι μια απλή προσέγγιση, καθώς ο κάθε κατασκευαστής μπορεί να προσδιορίσει τη μία και μόνη σωστή καταχώρηση. Επίσης, είναι σχετικά εύκολο να αποθηκεύσει αυτή την καταχώρηση στο ERP σύστημα της εταιρίας του. Ωστόσο, για τους πελάτες μπορεί να είναι δύσκολο να βρουν προϊόντα με κριτήριο ένα συγκεκριμένο σκοπό ή χρήση. Οι περισσότεροι αγοραστές ξέρουν για ποιο λόγο θα χρησιμοποιήσουν το προϊόν, αλλά μπορεί να μην ξέρουν τι σημαίνει αυτό σε σχέση με την απαιτούμενη ιδιότητα του προϊόντος.

Σύμφωνα με τις μετρικές που παρουσιάζονται στο [9] το πρότυπο UNSPSC παρουσιάζει τα παρακάτω χαρακτηριστικά :

(1) *Μέγεθος και ανάπτυξη* : Το UNSPSC πρότυπο ξεκίνησε με 19778 κλάσεις στην πρώτη του έκδοση και στην έκδοση 7,0901 έχει 20789 κλάσεις. Από το πλήθος των νέων και τροποποιημένων κλάσεων, το UNSPSC φαίνεται ότι αυξάνεται κατά 230 νέες κλάσεις ανά 30 ημέρες και επίσης φαίνεται ότι γίνεται σημαντική συντήρηση των υπαρχόντων καταχωρήσεων.

(2) *Ιεραρχία και ισορροπία περιεχομένου* : Το μεγαλύτερο μερίδιο κατηγοριών προέρχεται από πολύ λίγες διακλαδώσεις. Στο UNSPSC (έκδοση 7,0901) υπάρχει κυριαρχία λίγων κατηγοριών. Αναφορικά με τις κατηγορίες που αναπαριστούν υπηρεσίες, το UNSPSC περιλαμβάνει 4313 έννοιες υπηρεσιών, δηλαδή ποσοστό 21%. Αυτό δεν περιλαμβάνει υπηρεσίες που είναι "κρυμμένες" σε βαθύτερα επίπεδα της ιεραρχίας.

(3) *Βιβλιοθήκη ιδιοτήτων* : Το πρότυπο UNSPSC δεν περιλαμβάνει ιδιότητες προϊόντων.

4.2 Συμπεράσματα

Από την ανάλυση των δυο προτύπων συμπερασματικά, φαίνεται ότι και τα δυο οριζόντια πρότυπα περιλαμβάνουν έναν εντυπωσιακό αριθμό από κατηγορίες προϊόντων και υπηρεσιών , αλλά οι κατηγορίες είναι αρκετά άνισα κατανεμημένες μεταξύ των διαφόρων τμημάτων υψηλού επιπέδου. Οι ετικέτες και ο αριθμός των κατηγοριών υψηλού επιπέδου υπόσχονται μια πολύ ευρεία, ουδέτερης-βιομηχανίας πεδίο εφαρμογής, το οποίο είναι μια ανεκπλήρωτη αξίωση στο τρέχον στάδιο των προτύπων. Το eCl@ss και το UNSPSC είναι πολύ πιο ομοιόμορφα συμπληρωμένες, αλλά εξακολουθούν να έχουν 7 φορές (eCl@ss) και 11 φορές (UNSPSC) περισσότερες καταχωρήσεις στην μεγαλύτερή τους κατηγορία. Και οι δυο έχουν πάνω από 30% όλων των καταχωρήσεων στα τρία μεγαλύτερα τμήματα, και συνεπώς, μόνο μια μικρή διαμέριση των 25 (eCl@ss) , 55 (UNSPSC) κατηγοριών υψηλού επιπέδου.

Όταν βλέπουμε το πλήθος των άμεσων απογόνων για κάθε κόμβο, μπορούμε να δούμε ότι ο βαθμός πληρότητας μειώνεται στο eCl@ss από πάνω προς τα κάτω, η διακύμανση αυξάνεται από 61% (Υψηλότερο επίπεδο → $2^{ο}$ Επίπεδο) σε 156% ($3^{ο}$ Επίπεδο → $4^{ο}$ Επίπεδο), ενώ είναι πολύ πιο συνεπής στο UNSPSC (74% σε σύγκριση με το 100%). Και στα δυο πρότυπα ωστόσο, ο πληθυσμός στο επίπεδο των φύλλων ποικίλλει σε μεγάλο βαθμό, με ελάχιστο ένα μόνο φύλλο και μέγιστο 85 (eCl@ss) ή 92 (UNSPSC).

Φυσικά δεν μπορεί κανείς να υποθέσει ότι όλοι οι κλάδοι χρειάζονται τον ίδιο αριθμό εγγραφών, αλλά αυτό δεν δικαιολογεί την τάξη ή το μέγεθος που βρέθηκε στα τωρινά πρότυπα. Περιληπτικά, ο συνολικός αριθμός των κλάσεων “κρύβει” το γεγονός ότι πολλές από τις διακλαδώσεις εξακολουθούν να είναι πολύ ελλειπείς, και οι πιθανοί χρήστες πρέπει να ελέγξουν την κάλυψη που παρέχουν οι καταχωρήσεις στον τομέα τους πριν την υιοθέτηση κάποιου προτύπου.

Όσο αφορά τον βαθμό εξειδίκευσης, αυτός μπορεί να αξιολογηθεί καλύτερα με την εξέταση συγκεκριμένων αναθέσεων ιδιοτήτων σε κλάσεις. Εδώ το eCl@ss έχει ποσοστό μόνο 43%, ενώ το UNSPSC δεν έχει καθόλου ιδιότητες. Με άλλα λόγια, περισσότερο από τις μισές από όλες τις eCl@ss κλάσεις είναι επί του

παρόντος χωρίς συγκεκριμένες λίστες ιδιοτήτων. Από την άλλη μεριά, όλα τα πρότυπα περιέχουν πολλές ιδιότητες που χρησιμοποιούνται μόνο σε μια ή δυο κλάσεις. Αυτό δείχνει είτε ότι υπάρχει πλεονασμός, είτε “αυθαίρετη” δημιουργία λιστών προϊόντων κατ’απαίτηση, ή συνδυασμό των δυο.

Αναφορικά με την συντήρηση, και το eCl@ss και το UNSPSC υποβάλλονται σε συνεχή βελτίωση, με έναν μέσο όρο πάνω από 200 νέες τάξεις ανά μήνα.

Επί του παρόντος, υπάρχουν διαθέσιμα πολλά συστήματα και δεν αναμένεται ότι ένα μόνο πρότυπο ταξινόμησης θα χρησιμοποιείται σε όλο τον κόσμο και θα εκτείνεται σε όλες τις βιομηχανίες. Οι τωρινές περιγραφικές γλώσσες για προϊόντα και υπηρεσίες που περιγράφτηκαν δεν έχουν ούτε την απαιτούμενη κάλυψη των εννοιών ούτε την απαιτούμενη σημασιολογική ακρίβεια. Φαίνεται ότι η αντικατάσταση του ανθρώπινου στοιχείου σε ηλεκτρονικές διαδιακασίες αγοράς είναι πιο πολύπλοκη από ό,τι πολλοί ανέμεναν [45].

Η ανάλυση των προτύπων αυτών και η διαπίστωση της πολυπλοκότητας, του μεγάλου πλήθους των κλάσεων και των λοιπών μειωνεκτημάτων μας οδήγησαν στην υιοθέτηση μιας πιο ελαφριάς (lightweight) οντολογίας προϊόντων για τις ανάγκες του συστήματός μας, που περιγράφεται στην Ενότητα 5.2, όπου η εισαγωγή στιγμιοτύπων και κλάσεων είναι πρακτικά εφικτή και το μέγεθος της οντολογίας είναι μικρό, ώστε να επιτρέπεται η αυτόματη επεξεργασία της οντολογίας των προϊόντων. Μια τέτοια οντολογία θα μπορούσε να χρησιμοποιηθεί από ένα shopbot ώστε σε μικρό χρόνο να μπορεί να προτείνει προϊόντα τα οποία ταιριάζουν στις απαιτήσεις και τα ιδιαίτερα χαρακτηριστικά των χρηστών. Με αυτό τον τρόπο εκμεταλλευόμαστε τα πλεονεκτήματα που προσφέρουν τα shopbots όπως επίσης και τα πλεονεκτήματα της σημασιολογικής περιγραφής των προιόντων, χωρίς να χρειαστεί να καταφύγουμε σε χρονοβόρες και κοστοβόρες λύσεις.

ΚΕΦΑΛΑΙΟ 5

ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΚΑΙ ΣΧΕΔΙΑΣΜΟΣ ΣΥΣΤΗΜΑΤΟΣ

5.1 Γενική Αρχιτεκτονική του Συστήματος

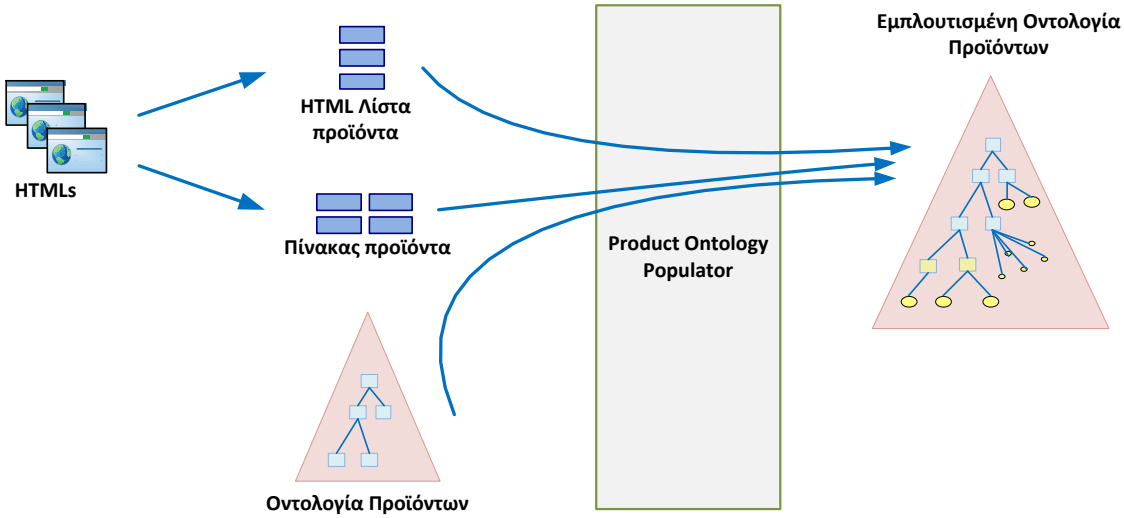
Σε μια εικονική αγορά λειτουργούν εκτός από τις οντότητες των αγοραστών και των πωλητών και ενδιάμεσες οντότητες, οι οποίες είναι βοηθητικές οντότητες με σκοπό τη διαχείριση, δηλαδή την διευκόλυνση, αλλά και την ασφαλή λειτουργία της αγοράς. Μια από τις ενδιάμεσες οντότητες είναι οι οντότητες αγορών (Shopping Bots – shopbots). Τα shopbots ή αυτόματες μηχανές αγοράς, είναι μια κατηγορία έξυπνων πρακτόρων (intelligent agents) που αναζητούν (αυτόματα και αποτελεσματικά), σε ένα μεγάλο αριθμό πωλητών, και παρέχουν στον καταναλωτή σχεδόν όλες τις πληροφορίες, σχετικά με ένα προϊόν, που υπάρχουν μια δεδομένη στιγμή στο διαδίκτυο. Η ύπαρξή τους προήλθε από την ανάγκη εύρεσης όλων των προϊόντων που συμπίπτουν με τις προτιμήσεις του καταναλωτή. Συνήθως η εύρεση του κατάλληλου αγαθού που επιθυμεί ο πελάτης είναι μια διαδικασία πολύ δύσκολη. Επίσης, είναι χρονοβόρα και πολλές φορές αναποτελεσματική λόγω της έλλειψης χρόνου ή προσοχής από τον χρήστη. Ο χρήστης μπορεί να βρει ένα μικρό ποσοστό των προσφορών που τον ενδιαφέρουν, ενώ με την χρήση των shopbots έχει πρόσβαση σε όλες τις προσφορές και μπορεί να επιλέξει αυτό που ταιριάζει περισσότερο στις προτιμήσεις του.

Η λειτουργία των shopbots μπορεί να βελτιστοποιηθεί με την χρήση οντολογιών γιατί πετυχαίνουμε μια πιο ευέλικτη απεικόνιση των δεδομένων, αλλά και των χαρακτηριστικών τους. Έτσι, η δήλωση των χαρακτηριστικών των προϊόντων μπορεί να γίνει πιο εξειδικευμένη και να παρουσιαστούν λιγότερα άχρηστα αποτελέσματα στον χρήστη μετά από κάθε αναζήτηση. Ακόμα, έχουμε μικρότερο χρόνο εκτέλεσης, αφού για την εύρεση ενός χαρακτηριστικού, δεν χρειάζεται αναζήτηση σε όλες τις σελίδες HTML των ηλεκτρονικών καταστημάτων, αλλά αρκεί ένα επερώτημα στην αντίστοιχη οντολογία.

Στην εργασία αυτή, δημιουργήθηκε ένας αλγόριθμος ο οποίος μπορεί να χρησιμοποιηθεί από ένα shopbot, ώστε να ανακτώνται προϊόντα και χαρακτηριστικά τους από πολλές HTML σελίδες και να αποθηκεύονται σε μια οντολογία προϊόντων. Με την χρήση της εμπλουτισμένης οντολογίας, που αποτελεί κατάλογος προϊόντων από διάφορες πηγές-πωλητές, αρκούν επερωτήσεις στην οντολογία με βάση συγκεκριμένα χαρακτηριστικά ώστε να επιστραφούν τα κατάλληλα αποτελέσματα στον χρήστη.

Ο καταναλωτής (είτε ανθρώπινη είτε αυτόματη οντότητα), παρουσιάζει μια περιγραφή των χαρακτηριστικών των προϊόντων που προτιμά (περιγραφή προϊόντος, αποδεκτό εύρος τιμής, τοποθεσία ηλεκτρονικού καταστήματος, τρόπος απόκτησης του αγαθού κτλ.). Το shopbot στην συνέχεια αναζητά στην εμπλουτισμένη οντολογία προϊόντων και αναλαμβάνει να βρει τα προϊόντα που αντιστοιχούν στην περιγραφή αυτή.

Τα συστατικά που συνθέτουν τη λογική αρχιτεκτονική του συστήματος απεικονίζονται στο Σχήμα 5.1. Στη συνέχεια δίνεται μια περιγραφή της λειτουργικότητας των συστατικών και ο γενικός σχεδιασμός του συστήματος.



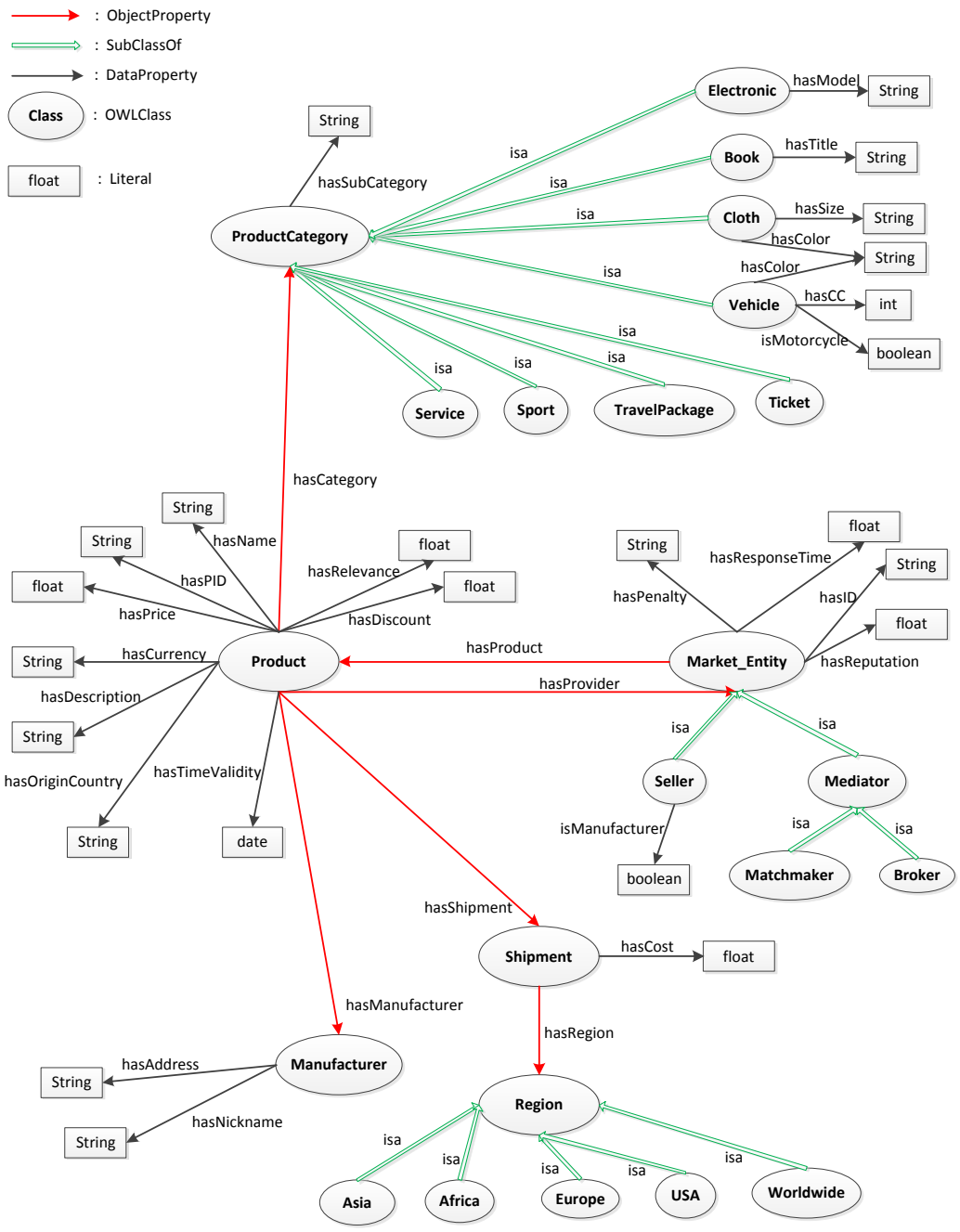
Σχήμα 5.1 : Η γενική αρχιτεκτονική του συστήματος

Ο σκοπός για τον οποίο σχεδιάστηκε και υλοποιήθηκε ο αλγόριθμος είναι ο εμπλουτισμός μιας υπάρχουσας οντολογίας προϊόντων με νέα στιγμιότυπα προϊόντων, τα οποία ανακτώνται από HTML σελίδες του Ιστού. Όπως είναι γνωστό, η πληθώρα των ηλεκτρονικών καταστημάτων που υπάρχουν στον Ιστό σήμερα, παρουσιάζουν τα προϊόντα τους σε HTML λίστες ή σε πίνακες. Αυτά αποτελούν την πηγή για άντληση μεγάλης ποσότητας πληροφορίας που αφορούν τα προϊόντα και τα χαρακτηριστικά τους. Σκοπός του συστήματός μας είναι αυτή η πληροφορία να αποθηκευτεί σε μια οντολογία, ώστε η εμπλουτισμένη με νέα στιγμιότυπα και κλάσεις τελική οντολογία να μπορεί να χρησιμοποιηθεί περαιτέρω για άλλες πιο πολύπλοκες εργασίες. Όπως φαίνεται και στην αρχιτεκτονική του (Σχήμα 5.1), το σύστημα έχει ως είσοδο τις HTML σελίδες, από τις οποίες θα ανακτήσει τα προϊόντα με τα χαρακτηριστικά τους, και ως δεύτερη είσοδο μια οντολογία προϊόντων. Η έξοδος του συστήματος είναι η εμπλουτισμένη οντολογία προϊόντων, με νέα στιγμιότυπα προϊόντων και νέες κλάσεις-κατηγορίες προϊόντων (αν δεν υπάρχουν ήδη στην αρχική οντολογία).

Ως στοιχείο εισόδου στο σύστημα χρησιμοποιήθηκαν οι HTML σελίδες, οι οποίες στην πράξη παρουσιάζουν προϊόντα που είναι αποθηκευμένα τοπικά σε σχεσιακές βάσεις των παρόχων των ηλεκτρονικών καταστημάτων. Η εξαγωγή των χαρακτηριστικών των προϊόντων απευθείας από τις σχεσιακές βάσεις των ηλεκτρονικών καταστημάτων δεν είναι εφικτή, γιατί δεν υπάρχει εύκολη πρόσβαση σε αυτές από εξωτερικές τρίτες οντότητες. Αντίθετα, οι HTML σελίδες των ηλεκτρονικών καταστημάτων βρίσκονται διάχυτες στον Ιστό και είναι εύκολα προσβάσιμες από όλους. Επιπλέον, οι HTML σελίδες παρουσιάζουν όλη την πληροφορία για χαρακτηριστικά των προϊόντων που θέλουν να μοιραστούν οι πωλητές με τους πιθανούς αγοραστές, “κρύβοντας” τις λεπτομέρειες της εσωτερικής υλοποίησης των σχεσιακών βάσεων, πράγμα που κάνει την διαδικασία της ανάκτησης προϊόντων πιο απλή και γρήγορη.

5.2 Η Οντολογία Προϊόντων

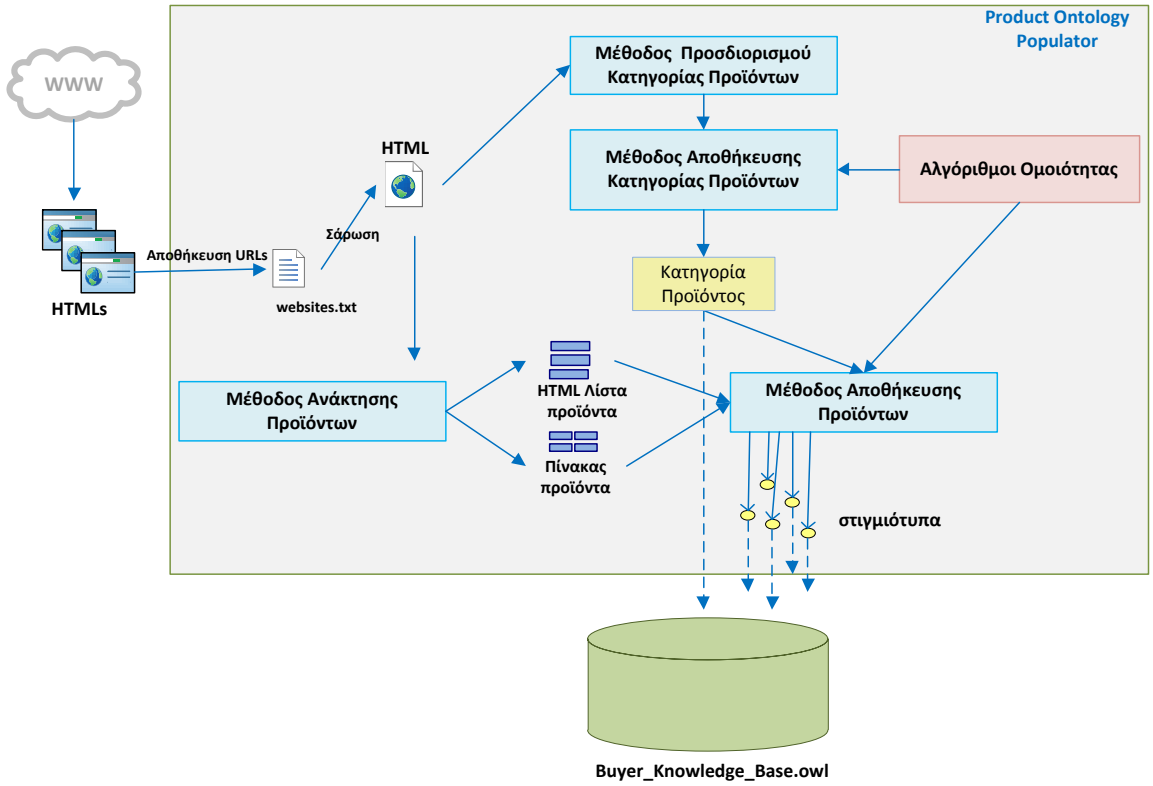
Ένα βασικό συστατικό στοιχείο του συστήματος αποτελεί η οντολογία προϊόντων, πάνω στην οποία θα προστεθούν τα νέα στιγμιότυπα προϊόντων. Στο Σχήμα 5.2 απεικονίζεται το εννοιολογικό σχήμα της οντολογίας “Buyer_Knowledge_Base.owl” που χρησιμοποιείται από το σύστημα. Χρησιμοποιώντας το Protégé 3.4.6 μπορεί κανείς να δει και να ενημερώσει χειροκίνητα την οντολογία. Όπως μπορούμε να δούμε, η οντολογία αποτελείται από τις βασικές κλάσεις “Product” που αναπαριστά το προϊόν, “ProductCategory” που αναπαριστά την κατηγορία προϊόντος, “Manufacturer” που αναπαριστά τον κατασκευαστή του προϊόντος, “Market_Entity” που αναπαριστά την αγορά (π.χ. ηλεκτρονικό κατάστημα) όπου εμφανίζεται το προϊόν, “Shipment” που αναπαριστά την αποστολή του προϊόντος στον αγοραστή και “Region” που αναπαριστά την τοποθεσία αποστολής προϊόντος. Οι κλάσεις αυτές συνδέονται μεταξύ τους με object-properties και κάθε κλάση έχει δικά της data-properties. Πρέπει να σημειωθεί ότι οι διάφορες κατηγορίες προϊόντων μοντελοποιούνται στην οντολογία σαν ξεχωριστές κλάσεις οι οποίες είναι υποκλάσεις της “ProductCategory”. Αυτό διευκολύνει την προσθήκη νέων κλάσεων που αναπαριστούν κατηγορίες προϊόντων και δεν υπάρχουν ήδη στην οντολογία. Έτσι, το σύστημά μας δημιουργεί νέες κλάσεις – κατηγορίες προϊόντων όταν δεν υπάρχει στην οντολογία η κατηγορία προϊόντων μιας σελίδας Ιστού. Η προσθήκη των νέων κλάσεων δεν συνοδεύεται από παράλληλη προσθήκη data-properties που μπορεί να ανήκουν στην νέα κατηγορία προϊόντων, γιατί αυτό απαιτεί πρόσθετο σχεδιασμό. Η οντολογία “Buyer_Knowledge_Base.owl” είναι ευέλικτη και μπορεί να χρησιμοποιηθεί για την περιγραφή των υλικών αγαθών και βασικών υπηρεσιών στο διαδίκτυο. Μπορεί να κλιμακωθεί (scale) πολύ καλά στις βασικές υποδομές του Σημασιολογικού Ιστού που είναι διαθέσιμες σήμερα. Κατά συνέπεια, μπορούμε εύκολα να εμπλουτίσουμε την οντολογία με νέα στιγμιότυπα και κλάσεις (όταν απαιτείται).



Σχήμα 5.2: Εννοιολογικό σχήμα της οντολογίας προϊόντων Buyer_Knowledge_Base.owl

5.3 Η Μεθοδολογία Δημιουργίας Προϊόντων στην Οντολογία

Η μεθοδολογία που ακολουθήσαμε για τη δημιουργία της εφαρμογής “Product Ontology Population” με σκοπό την δημιουργία στιγμιοτύπων προϊόντων από πίνακες και HTML λίστες σελίδων Ιστού απεικονίζεται στο Σχήμα 5.3.



Σχήμα 5.3: Μεθοδολογία δημιουργίας προϊόντων στην οντολογία

Ο αλγόριθμος της εφαρμογής Product Ontology Populator παρουσιάζεται στον παρακάτω Αλγόριθμο 1.

Αλγόριθμος 1. Ο αλγόριθμος Product Ontology Populator

```
For each strLine in "websites.txt"
{
    If (strLine!=Empty )
    {
       /*  ανέκτησε τον πάροχο ιστοσελίδας, π.χ. bestbuy.com  */
       Provider = getProvider(strLine)
```

```
        /*  ανέκτησε την κατηγορία προϊόντος της ιστοσελίδας  */
        product_category  = getProductCategory(strLine)

        If (product_category != Empty)
        {
          /* αποθήκευσε κατηγορία προϊόντος σε κλάση στην οντολογία και
             δημιούργησε στιγμιότυπο αυτού αν δεν υπάρχει   */
          product_category_instance = addProductCategory2Ontology(ontology,
                                      iri, manager,file, factory, pm, product_category)

           /* η μέθοδος εξάγει από την ιστοσελίδα "strLine" τα προϊόντα και
              φτιάχνει στιγμιότυπο τύπου Product στην οντολογία με
              Provider = range(hasProvider), product_category_instance  =
              range(hasCategory). Μετά βάζει τιμές στις λοιπές ιδιότητες
              προϊόντος στην οντολογία    */
           addProductProperties2Ontology(ontology, iri, Seller , Provider,
                                         manager, strLine, Product, file, factory, pm,
                                                   product_category_instance)

        }
      }
}

addProductProperties2Ontology(ontology, iri, Seller , Provider,
                              manager, strLine, Product, file, factory, pm,
                              product_category_instance)
/* η μέθοδος εξάγει από την ιστοσελίδα "strLine" τα προϊόντα και φτιάχνει
στιγμιότυπο τύπου Product στην οντολογία με  Provider  = range(hasProvider),
product_category_instance  = range(hasCategory). Μετά βάζει τιμές στις λοιπές
ιδιότητες προϊόντος στην οντολογία    */

{
   /* αποθήκευσε τον πάροχο ιστοσελίδας στην οντολογία */
  ProviderInstance = addProvider2Ontology(ontology, Seller, Provider, factory,
                                          manager, pm,file)
  For each ( <li>, <td>, <table> contains {"product", "prd", "item","result
            catalog","odd","even"})
  {
     /*  αποθήκευσε τις ιδιότητες προϊόντων της ιστοσελίδας στην οντολογία   */
     populateOntology(ontology, pm, manager, factory, file,
                      datapropertiesNames, product_category_instance,
                      Product, ProviderInstance, (TagNode "li" or "table" or "td"))
  }
}
```

populateOntology (ontology, pm, manager, factory, file, datapropertiesNames, product_category_instance, Product,ProviderInstance, TagNode “li” or “table” or “td”)

```
/*  Η μέθοδος αποθηκεύει στην οντολογία 1 στιγμιότυπο προϊόντος με τις μαζί με
    τις τιμές των ιδιοτήτων του που εξήχθησαν από την ιστοσελίδα */
{
    /*  δημιούργησε το στιγμιότυπο προϊόντος τύπου Product  */
    instance = createProductInstance(ontology, Product, manager, factory, pm,
                                          id_value)
    If (instance != Empty)
    {
        /* αποθήκευσε τιμή του object-property “hasCategory” για το στιγμιότυπο
           προϊόντος  */
       createObjectPropertyAssertion(hasCategory, instance,
                          product_category_instance, factory, ontology, manager, pm)
        /* αποθήκευσε τιμή του object-property “hasProvider” για το στιγμιότυπο
           προϊόντος  */
       createObjectPropertyAssertion(hasProvider, instance, ProviderInstance,
                                                    factory, ontology, manager, pm)
        /* ανέκτησε όλα τα “class” attributes του προϊόντος στην ιστοσελίδα  */
       attributeSet = getElementsHavingAttribute("class")
       If (attributeSet != Empty)
       {
           For each attribute A ϵ  attributeSet do
           {
              /* αποθήκευσε τις τιμές των “class” attributes στις κατάλληλες
                 ιδιότητες προϊόντων στην οντολογία   */
             getClassAttributesFromWebSite(datapropertiesNames,
                                    product_characteristic,characteristic_value,
                                     instance,ontology,factory,pm, manager)
           }
       }
        /*  έλεγξε αν υπάρχει στην οντολογία το προϊόν που μόλις δημιουργήθηκε.
           Αν υπάρχει διέραψέ το. */
      deleteDuplicateInstances(ontology, manager, factory, pm, file,
                                     Product, instance)
      /* Αποθήκευσε τις αλλαγές στην οντολογία  */
      saveOntology(manager, ontology, file)
    }
}
```

Με λίγα λόγια αυτό που κάνει η εφαρμογή είναι το εξής. Το πρώτο βήμα αφορά την εύρεση των ιστοσελίδων που περιέχουν προϊόντα τα οποία εμφανίζονται σε HTML λίστες ή πίνακες. Στη συνέχεια, αφού βρεθούν οι ιστοσελίδες, αποθηκεύουμε τα URLs τους στο αρχείο “websites.txt”. Το αρχείο “websites.txt” χρησιμοποιείται από το σύστημα γιατί περιέχει τα URLs των σελίδων Ιστού από τα οποία θέλουμε να ανακτήσουμε τα προϊόντα τους και να τα εισάγουμε στην οντολογία μας, την “Buyer_Knowledge_Base.owl”. Το “websites.txt” περιλαμβάνει ένα URL ανά γραμμή, έτσι ώστε αποθηκεύοντας πέντε ή δέκα URLs να γίνεται σάρωση των σελίδων αυτών μία προς μία σειριακά και αυτόματα, χωρίς να απαιτείται διάδραση με τον χρήστη.

Ανακτάται το URL μιας ιστοσελίδας που είναι αποθηκευμένη στο αρχείο “websites.txt”, σαρώνεται και “καθαρίζεται” η ιστοσελίδα με την χρήση του HTMLCleaner parser, ώστε να δημιουργηθεί ένα καλά δομημένο HTML έγγραφο, χωρίς να υπάρχουν tags που δεν κλείνουν κτλ. Στη συνέχεια, ανακτάται από την ιστοσελίδα η κατηγορία στην οποία ανήκουν τα προϊόντα. Αν αυτή η κατηγορία υπάρχει ήδη σαν κλάση στην οντολογία τότε σε αυτήν θα τοποθετηθούν και τα νέα προϊόντα που θα εισαχθούν, αν όμως η κατηγορία δεν υπάρχει στην οντολογία, τότε δημιουργείται νέα κατηγορία σαν υποκλάση της κλάσης “ProductCategory”. Αφού προστεθεί η κατηγορία προϊόντων, μετά ανακτώνται ένα ένα τα προϊόντα από τη λίστα ή τον πίνακα προϊόντων της ιστοσελίδας μαζί με τα χαρακτηριστικά τους. Δημιουργούνται νέα στιγμιότυπα προϊόντων στην οντολογία “Buyer_Knowledge_Base.owl” και αν βρεθεί λεξικογραφική ομοιότητα μεταξύ κάποιου χαρακτηριστικού προϊόντος στην ιστοσελίδα με κάποιο data-property στην οντολογία, τότε η τιμή του χαρακτηριστικού στην ιστοσελίδα προστίθεται ως τιμή του αντίστοιχου data-property στην οντολογία. Αυτή η διαδικασία εκτελείται επαναληπτικά για κάθε ιστοσελίδα που είναι αποθηκευμένη στο αρχείο “websites.txt”. Ως αποτέλεσμα, η οντολογία “Buyer_Knowledge_Base.owl” θα έχει εμπλουτιστεί με νέες κλάσεις που αναπαριστούν τις κατηγορίες προϊόντων, και με νέα στιγμιότυπα που περιέχουν τα προϊόντα και τα χαρακτηριστικά τους που ανακτήθηκαν από τις ιστοσελίδες.

Στις επόμενες παραγράφους γίνεται λεπτομερή περιγραφή των μεθόδων που απεικονίζονται στο Σχήμα 5.3.

5.3.1 Αλγόριθμοι Ομοιότητας

Οι αλγόριθμοι ομοιότητας που χρησιμοποιούνται στο σύστημα είναι δυο κατηγοριών :

- *Αλγόριθμοι σημασιολογικής ομοιότητας* : Ως σημασιολογική ομοιότητα ορίζουμε μια μετρική η οποία καθορίζει τα κοινά και τα διαφορετικά στοιχεία μεταξύ δυο εννοιών. Ως ποσοτική μέτρηση της σημασιολογικής ομοιότητας αποδίδεται συνήθως ένας αριθμός που υποδηλώνει το πόσο όμοιες ή όχι είναι δυο οντότητες. Για την υλοποίησή τους, οι αλγόριθμοι που χρησιμοποιούνται στο σύστημα χρησιμοποιούν ένα λεξικό που έχει τη μορφή ταξινομίας, το WordNet [21], ώστε να μετρηθεί η απόσταση που έχουν οι έννοιες μεταξύ τους και να βγει συμπέρασμα για τη συσχέτισή τους σημασιολογικά και όχι λεξικογραφικά. Συγκεκριμένα, οι αλγόριθμοι σημασιολογικής ομοιότητας που χρησιμοποιήθηκαν είναι : Αλγόριθμος Resnik [47], Αλγόριθμος Leacock-Chodorow [48], Αλγόριθμος Jiang-Conrath [49], Αλγόριθμος Lin [50], Αλγόριθμος Wu-Palmer [51], Αλγόριθμος Tversky [52], Μετρικές Li-Zuhair-Bandar-McLean [54], Αλγόριθμος Aggire-Rigau [55][56], Μετρική Μήκους Μονοπατιού , Αλγόριθμος Rada [53].

  Στο σύστημά μας, για την εύρεση σημασιολογικής ομοιότητας μεταξύ δυο εννοιών, χρησιμοποιείται ο μέσος όρος της τιμής που επιστρέφεται από όλους τους παραπάνω αλγορίθμους. Σχετικά με το πότε εκτελείται σημασιολογική ομοιότητα, πρέπει να αναφερθεί ότι αυτό συμβαίνει όταν γίνεται σύγκριση μεταξύ της κατηγορίας προϊόντος στην ιστοσελίδα, με το όνομα της κάθε κλάσης-κατηγορίας προϊόντος στην οντολογία, ώστε να προκύψει, μέσω του βαθμού σημασιολογικής ομοιότητας, είτε ότι η κατηγορία προϊόντος υπάρχει σαν κλάση στην οντολογία (βαθμός ομοιότητας >= 0.7) είτε ότι δεν υπάρχει (βαθμός ομοιότητας <0.7). Ο βαθμός ομοιότητας 0.7 προέκυψε από δοκιμές που έγιναν και φάνηκε ότι αυτή η τιμή φέρνει τα

καλύτερα αποτέλεσματα. Η χρήση όμως των αλγορίθμων σημασιολογικής ομοιότητας επιφέρει μεγάλη χρονική καθυστέρηση στην εκτέλεση της εφαρμογής, λόγω του μεγάλου όγκου και των πολλών συγκρίσεων που πραγματοποιούνται στο WordNet. Για τον λόγο αυτό, δεν χρησιμοποιήθηκε η σημασιολογική ομοιότητα σε άλλα σημεία της εφαρμογής που ίσως να ήταν χρήσιμο, όπως για παράδειγμα στη σύγκριση των ονομάτων ιδιοτήτων των προϊόντων στην ιστοσελίδα με τα ονόματα των ιδιοτήτων στην οντολογία. Κάτι τέτοιο θα επέφερε μη επιτρεπτές καθυστερήσεις όπως θα διαπιστώσουμε στα πειράματα αξιολόγησης. Πρέπει τέλος να σημειωθεί, ότι για να εκτελεστεί σημασιολογική ομοιότητα μεταξύ δυο εννοιών, αυτές πρέπει να είναι έγκυρες λέξεις στο WordNet. Αυτό στις περισσότερες περιπτώσεις δεν συμβαίνει, διότι οι κατηγορίες προϊόντων στις σελίδες Ιστού, είτε αποτελούνται από δυο (π.χ. Wedding Flowers), τρεις (π.χ. Small Kitchen Appliances) ή πιο σπάνια τέσσερις λέξεις. Σαν αποτέλεσμα αυτής της σχεδιαστικής πρακτικής των σελίδων Ιστού, τελικά για την εύρεση της ομοιότητας μεταξύ των κατηγοριών προϊόντων στον Ιστό με αυτές στην οντολογία, στις περισσότερες περιπτώσεις καταλήγει να χρησιμοποιείται η λεξικογραφική ομοιότητα που περιγράφεται στη συνέχεια.

- *Αλγόριθμοι λεξικογραφικής ομοιότητας* : Η λεξικογραφική ομοιότητα μεταξύ δυο συμβολοσειρών αποσκοπεί στην εξαγωγή αριθμητικής τιμής η οποία να υποδηλώνει τον βαθμό ομοιότητας μεταξύ τους. Υπάρχουν πολλοί αλγόριθμοι για την εύρεση λεξικογραφικής ομοιότητας. Στο σύστημά μας, η εύρεση της λεξικογραφικής ομοιότητας υπολογίζεται ως η μέση τιμή της ομοιότητας που επιστρέφουν οι αλγόριθμοι : q-gram [28], Lin [29], Jaro [30], Needleman-Wunch [31]. Λεξικογραφική ομοιότητα εκτελείται στο σύστημα κατά την σύγκριση των κατηγοριών προϊόντων στην οντολογία με αυτή της σελίδας Ιστού για να βρεθεί το καλύτερο ταίριασμα. Στο σημείο αυτό εκτελείται λεξικογραφική ομοιότητα μόνο αν η σημασιολογική ομοιότητα (που περιγράφηκε νωρίτερα) δεν εφαρμοστεί, γιατί κάποια από τις δυο συμβολοσειρές εισόδου δεν αποτελεί έγκυρη λέξη στο WordNet. Επίσης, λεξικογραφική ομοιότητα χρησιμοποιείται και κατά την προσθήκη τιμών

ιδιοτήτων στα στιγμιότυπα προϊόντων στην οντολογία. Γενικά στο σύστημα όπου εκτελείται σύγκριση για εύρεση ομοιότητας μεταξύ δυο συμβολοσειρών, πραγματοποιείται λεξικογραφική ομοιότητα, εκτός από το σημείο της σύγκρισης για την κατηγορία προϊόντος, όπου εκτελείται σημασιολογική ομοιότητα αν οι λέξεις εισόδου υπάρχουν στο WordNet. Οι αλγόριθμοι λεξικογραφικής ομοιότητας είναι γρήγοροι στην εκτέλεση και δεν επιφέρουν καθυστερήσεις στο σύστημα. Επιπλέον, έχουν αρκετά καλά αποτελέσματα όσον αφορά στη σύγκριση των ιδιοτήτων των προϊόντων στην οντολογία με αυτές στις σελίδες Ιστού.

5.3.2 Μέθοδος Προσδιορισμού Κατηγορίας Προϊόντων

Η Μέθοδος Προσδιορισμού Κατηγορίας Προϊόντων είναι η πρώτη που εκτελείται στο σύστημά μας και αφορά στον προσδιορισμό της κατηγορίας στην οποία ανήκουν τα προϊόντα που εμφανίζονται σε μια σελίδα Ιστού. Είναι σημαντικό να ανακτηθεί η σωστή κατηγορία προϊόντων, γιατί κάτω από αυτή θα “κρεμαστούν” τα στιγμιότυπα προϊόντων οντολογίας που θα δημιουργηθούν σε επόμενο στάδιο. Από μελέτη πολλών ηλεκτρονικών καταστημάτων στον Ιστό, ώστε να βρεθεί ένα κοινό πρότυπο εμφάνισης της κατηγορίας προϊόντων, διαπιστώθηκε ότι αυτό συμβαίνει με την χρήση των html headers (h1,h2,h3 κτλ.) και <title> στοιχείων. Στις πλείστες των περιπτώσεων το <h1> φέρνει την σωστή κατηγορία, ενώ τα h2,h3 κτλ. περιλαμβάνουν επιμέρους ονόματα προϊόντων που εμφανίζονται στην ιστοσελίδα. Το <title> επίσης περιέχει την κατηγορία προϊόντος, αλλά ταυτόχρονα εμπεριέχει και άλλη πληροφορία (π.χ. <title> Tennis - Women's and Men's Designer Tennis Clothes & Accessories Online at harrods.com </title>) και δεν είναι πάντα εύκολο να εξαχθεί από το κείμενο μόνο η κατηγορία, γιατί κάθε ιστοσελίδα μπορεί να έχει διαφορετικό τρόπο καταγραφής του <title>. Επειδή παρατηρήθηκε ότι συνήθως στο <title> η κατηγορία προϊόντος εμφανίζεται πριν τα ειδικά σύμβολα { “|”, “-“ , “:” “,” }. Έτσι, στον αλγόριθμό μας, εξάγονται οι λέξεις που υπάρχουν πριν αυτά τα σύμβολα (π.χ. Tennis) και αυτές θεωρούνται ως κατηγορία προϊόντων. Αν η κατηγορία

προϊόντων που εξάγεται από το <title> είναι πάνω από 4 λέξεις, κρατούνται μόνο οι 4 πρώτες λέξεις. Τελικά, με μια σύγκριση του πλήθους των λέξεων από τις οποίες αποτελούνται τα <h1> και <title>, επιλέγεται ως κατηγορία προϊόντων αυτό με τις λιγότερες λέξεις, δίνοντας προτεραιότητα στο <h1> αν έχουν ίδιο αριθμό λέξεων, γιατί το <title> είναι πιο “φλύαρο” όπως είπαμε από ότι το <h1>. Επειδή στα ονόματα κλάσεων στην οντολογία δεν μπορούν να υπάρχουν κλάσεις που να περιέχουν ειδικά σύμβολα {“&”, “,”} και το κενό, γι αυτό όλα αντικαθίστανται με “_” (π.χ. κατηγορία “Christmas Food” θα αποθηκευτεί ως “Christmas_Food”). Πρέπει τέλος να σημειωθεί ότι τα αποτελέσματα προσδιορισμού της κατηγορίας προϊόντων με αυτή τη μέθοδο είναι αρκετά καλά, όπως φαίνεται και στα σενάρια εκτέλεσης στο επόμενο κεφάλαιο. Επιπλέον, ο αλγόριθμος λειτουργεί ικανοποιητικά όταν τα <h2>, <h3> ή <h4> χρησιμοποιούνται για τον προσδιορισμό της κατηγορίας προϊόντος.

Τα βήματα εκτέλεσης του αλγορίθμου περιγράφονται αναλυτικά ως εξής:

1. Διαβάζεται το αρχείο “websites.txt” και ανακτάται η πρώτη γραμμή που περιέχει το URL κάποιας ιστοσελίδας. Στη συνέχεια, ο αλγόριθμος επισκέπτεται και σαρώνει την ιστοσελίδα, ώστε να βρεθεί η κατηγορία στην οποία ανήκουν τα προϊόντα που εμφανίζονται.
2. Ανακτάται η τιμή που έχει το στοιχείο <title>. Συχνά η τιμή του περιλαμβάνει μια πρόταση που είναι χωρισμένη με κάποιον από τους χαρακτήρες {"|","-",":"}. Αν συμβαίνει αυτό αποθηκεύεται μόνο η συμβολοσειρά που υπάρχει πριν από κάποιον από αυτούς τους χαρακτήρες. Διαφορετικά, αν δεν υπάρχει τέτοιος χαρακτήρας στην τιμή του <title> αποθηκεύεται ολόκληρη η τιμή του. Αν το <title> αποτελείται από περισσότερες από 4 λέξεις, αποθηκεύονται μόνο οι 4 πρώτες, με την λογική ότι δεν μπορεί να υπάρξει κατηγορία προϊόντος που να αποτελείται από περισσότερες από 4 λέξεις.
3. Ανακτώνται οι τιμές που έχουν τα header tags.
4. Γίνεται σύγκριση του πλήθους των λέξεων από τις οποίες αποτελείται το <title> και το ελάχιστο πλήθος των header tags. Όποιο από τα δυο έχει τις λιγότερες λέξεις αυτό αποθηκεύεται σαν κατηγορία των προϊόντων της

ιστοσελίδας. Αυτό γίνεται, γιατί όπως είναι γνωστό, τα ηλεκτρονικά καταστήματα προσπαθούν να παρουσιάσουν τις κατηγορίες των προϊόντων όσο πιο συνοπτικά γίνεται, με λίγες λέξεις που περιγράφουν επακριβώς την κατηγορία του προϊόντος. Τα στοιχεία (tags) που περιλαμβάνουν πολλές λέξεις συνήθως περιέχουν πρόσθετη πληροφορία, πέρα από την κατηγορία προϊόντος, την οποία δεν θέλουμε. Έτσι αποθηκεύεται ως κατηγορία προϊόντος το στοιχείο με τις λιγότερες λέξεις.

5. Όταν η κατηγορία προϊόντος αποτελείται από περισσότερες από δυο λέξεις (π.χ. “Cooker Hoods”) τότε αυτές ενώνονται σε μια (δηλαδή “Cooker_Hoods”) ώστε να μπορεί να αποθηκευτεί σαν κλάση στην οντολογία σε επόμενο στάδιο.

5.3.3 Μέθοδος Αποθήκευσης Κατηγορίας Προϊόντων

Επόμενο στάδιο της μεθοδολογίας υλοποίησης που αναπτύχθηκε είναι η αποθήκευση της κατηγορίας προϊόντων στην οντολογία. Η οντολογία προϊόντων “Buyer_Knowledge_Base.owl” είναι το δεύτερο στοιχείο εισόδου στο σύστημα (μετά το αρχείο “websites.txt”) και παρουσιάστηκε αναλυτικά στην Ενότητα 5.2. Η οντολογία χρησιμοποιείται ως βάση από το σύστημα, ώστε να εισαχθούν νέα στιγμιότυπα και κλάσεις (όπου χρειάζεται). Τα στιγμιότυπα που θα εισαχθούν προέρχονται από τα προϊόντα που υπάρχουν στις ιστοσελίδες που έχουν δοθεί στο αρχείο “websites.txt”.

Η Μέθοδος Αποθήκευσης Κατηγορίας Προϊόντων αποτελείται από τα παρακάτω βήματα :

1. Φορτώνεται η οντολογία “Buyer_Knowledge_Base.owl” που είναι αποθηκευμένη στο υπολογιστικό σύστημα.
2. Ανακτώνται όλες οι υποκλάσεις της κλάσης “ProductCategory” και υπολογίζεται η ομοιότητα μεταξύ κάθε υποκλάσης του “ProductCategory” με την κατηγορία προϊόντων στην ιστοσελίδα. Για τον υπολογισμό της ομοιότητας χρησιμοποιήθηκαν οι αλγόριθμοι που περιγράφονται στην

Ενότητα 5.3.1 παραπάνω. Συγκεκριμένα, αν η κατηγορία προϊόντων στην ιστοσελίδα αποτελεί έγκυρη λέξη στο WordNet, τότε εκτελούνται οι αλγόριθμοι σημασιολογικής ομοιότητας, αν όχι εκτελούνται οι αλγόριθμοι λεξικογραφικής ομοιότητας. Τελικά, από το αποτέλεσμα της ομοιότητας εντοπίζεται η υποκλάση του “ProductCategory” που έχει την μέγιστη ομοιότητα με την κατηγορία προϊόντων στην ιστοσελίδα.

- Αν η ομοιότητα είναι μεγαλύτερη ή ίση από 0.7, τότε υπολογίζεται η λεξικογραφική ομοιότητα κάθε στιγμιοτύπου της κλάσης-κατηγορίας προϊόντος (που εντοπίστηκε με την μεγαλύτερη ομοιότητα) με την κατηγορία προϊόντος της ιστοσελίδας. Αυτό γίνεται για να δούμε αν η κατηγορία προϊόντων στην ιστοσελίδα υπάρχει ήδη σαν στιγμιότυπο κάποιας υποκλάσης του “ProductCategory” στην οντολογία. Αν υπάρχει, τότε αυτό το στιγμιότυπο αποθηκεύεται για να χρησιμοποιηθεί στην “*Μέθοδο Αποθήκευσης Προϊόντων*”. Διαφορετικά, δημιουργείται νέο στιγμιότυπο της κλάσης-κατηγορίας προϊόντος που υπάρχει στην οντολογία. Το όνομα του νέου στιγμιότυπου θα είναι το “όνομα της κατηγορίας συν ένας τυχαίος αριθμός στο διάστημα [1...100]”. Αυτό γίνεται για να αποφευχθεί η δημιουργία ίδιων ονομάτων στιγμιοτύπων στην οντολογία.
- Διαφορετικά, δημιουργείται μια νέα κλάση, που θα είναι υποκλάση του “ProductCategory”. Η νέα κλάση θα έχει ως όνομα την κατηγορία προϊόντων που ανακτήθηκε από το προηγούμενο βήμα “*Μέθοδος Προσδιορισμού Κατηγορίας Προϊόντων*”. Επίσης, δημιουργείται και ένα στιγμιότυπο της νέας κλάσης-κατηγορίας προϊόντων, που θα έχει ως όνομα, το “όνομα κατηγορίας προϊόντος συν έναν τυχαίο αριθμό στο διάστημα [1...100]”. Ο τρόπος αυτός δημιουργίας ονομασίας επιλέχθηκε για να μην δημιουργηθεί κλάση και στιγμιότυπο με το ίδιο όνομα, γιατί τότε η οντολογία που θα προκύψει δεν θα είναι συντακτικά ορθή.

- Τελικά, αν δημιουργήθηκε νέα κλάση που θα είναι η κατηγορία προϊόντων και το αντίστοιχο στιγμιότυπό της αποθηκεύονται στην οντολογία “Buyer_Knowledge_Base.owl”.

5.3.4 Μέθοδος Ανάκτησης Προϊόντων

Το επόμενο βήμα της μεθοδολογίας αφορά τον τρόπο εντοπισμού των προϊόντων που εμφανίζονται σε μια ιστοσελίδα, καθώς και τον τρόπο ανάκτησης των χαρακτηριστικών τους. Επειδή η εργασία βασίζεται σε ιστοσελίδες που εμφανίζουν τα προϊόντα σε HTML λίστες και πίνακες, το βήμα αυτό περιλαμβάνει εντοπισμό των λιστών που περιλαμβάνουν προϊόντα και τον εντοπισμό των πινάκων που περιέχουν προϊόντα. Αναλυτικά τα βήματα του αλγορίθμου έχουν ως ακολούθως:

1. Αρχικά, σαρώνεται η ιστοσελίδα με τη χρήση του HTMLCleaner parser, και ανακτώνται όλα τα <li> και τα <table>, <td> στοιχεία (elements).
2. Από τα στοιχεία λιστών και πίνακα που ανακτήθηκαν, εντοπίζονται αυτά που αφορούν παρουσίαση προϊόντος, χρησιμοποιώντας συγκεκριμένες λέξεις αναγνώρισης όπως {“product”, “prd”, “item”, “result catalog”, “odd”, “even”}. Η λίστα αυτή είναι προτεινόμενη και έχει εξαχθεί μετά από πειράματα που έγιναν σε διάφορες ιστοσελίδες. Φυσικά, μπορεί να επεκταθεί πολύ εύκολα ώστε να περιλάβει νέες λέξεις αναγνώρισης προιόντων. Τα υπόλοιπα στοιχεία λιστών και πινάκων που τυχόν υπάρχουν σε μια HTML σελίδα αγνοούνται, γιατί θεωρούμε ότι χρησιμοποιούνται για λόγους καλύτερης παρουσίασης των λοιπών περιεχομένων των HTML σελίδων και δεν περιλαμβάνουν περιγραφή προϊόντων.
3. Για κάθε στοιχείο προϊόντος που εντοπίστηκε στο προηγούμενο βήμα, ανακτώνται όλα τα εμφωλευμένα στοιχεία που περιέχουν την περιγραφή του προϊόντος με τα χαρακτηριστικά του.

4. Το προϊόν μαζί με τα χαρακτηριστικά που ανακτήθηκαν χρησιμοποιούνται στη συνέχεια από την "*Μέθοδο Αποθήκευσης Προϊόντων*", όπως περιγράφεται παρακάτω.

5.3.5 Μέθοδος Αποθήκευσης Προϊόντων

Το τελικό στάδιο της μεθοδολογίας υλοποίησης του συστήματος αποτελεί η αποθήκευση στην δοσμένη οντολογία "Buyer_Knowledge_Base.owl" των προϊόντων που βρέθηκαν στους πίνακες και στις λίστες προϊόντων μιας ιστοσελίδας. Από το προηγούμενο στάδιο έχουν εξαχθεί τα προϊόντα (δηλαδή τα εν δυνάμει στιγμιότυπα προϊόντων στην οντολογία) με τα χαρακτηριστικά τους. Τα βήματα που εκτελούνται για την αποθήκευση των προϊόντων στην οντολογία είναι τα εξής:

1. Για κάθε προϊόν που ανακτήθηκε δημιουργείται ένα στιγμιότυπο τύπου "Product" στην οντολογία, το οποίο θα έχει όνομα "prod συν τυχαίο long αριθμό". Η παραγωγή του τυχαίου long αριθμού γίνεται με έναν randomGenerator της κλάσης Random της Java. Αυτό γίνεται ώστε να βεβαιώνεται η μοναδικότητα κάθε νέου στιγμιότυπου προϊόντος που δημιουργείται και να μην τύχει να δημιουργηθούν δυο διαφορετικά προϊόντα με το ίδιο όνομα, που θα έχει σαν αποτέλεσμα να δημιουργηθεί ένα στιγμιότυπο στην οντολογία, το οποίο θα έχει χαρακτηριστικά και των 2 προϊόντων. Επίσης, πριν δημιουργηθεί κάθε νέο στιγμιότυπο ελέγχεται ότι δεν υπάρχει ήδη στιγμιότυπο με το ίδιο όνομα στην οντολογία.
2. Επόμενο βήμα είναι η δημιουργία και αποθήκευση στην οντολογία στιγμιότυπου του παρόχου προϊόντος. Ο πάροχος προϊόντος αναπαρίσταται στην οντολογία με την κλάση "Seller". Ανακτάται από την ιστοσελίδα ο πάροχος (π.χ. skroutz.gr, bestbuy.com κτλ.) και υπολογίζεται η λεξικογραφική ομοιότητα (όπως περιγράφτηκε στην ενότητα 5.3.1 : Αλγόριθμοι Ομοιότητας) μεταξύ κάθε στιγμιοτύπου της κλάσης "Seller" με το όνομα του παρόχου στην ιστοσελίδα.

- Αν η ομοιότητα είναι μεγαλύτερη ή ίση με 0.9, τότε ο πάροχος υπάρχει ήδη σαν στιγμιότυπο στην οντολογία. Η τιμή εδώ είναι 0.9 και όχι 1, που σημαίνει πλήρης ταύτιση.
- Διαφορετικά, δημιουργείται νέο στιγμιότυπο της κλάσης “Seller” με το όνομα του παρόχου ιστοσελίδας.

3. Για κάθε στιγμιότυπο τύπου “Product” που προστίθεται στην οντολογία δημιουργούνται τα object-property assertions “hasCategory” , “hasProvider” τα οποία έχουν domain, το στιγμιότυπο προϊόντος. Το “hasCategory” θα έχει range το στιγμιότυπο της κατηγορίας που δημιουργήθηκε νωρίτερα (στην “*Μέθοδο Αποθήκευσης Κατηγορίας Προϊόντων*”). Το “hasProvider” θα έχει range το στιγμιότυπο της κλάσης “Seller” που μόλις δημιουργήθηκε (βήμα 2).
4. Τα html στοιχεία που περιλαμβάνουν ένα προϊόν (όπως ανακτήθηκαν μέσω του “*Αλγορίθμου Ανάκτησης Προϊόντων*”) έχουν εμφωλευμένα πολλά στοιχεία τα οποία χαρακτηρίζονται με το attribute “class” και αποτελούν τα χαρακτηριστικά του προϊόντος. Αυτό που κάνει το σύστημα είναι να παίρνει όλα τα “class” attributes και τις τιμές τους και να τις συγκρίνει με τα ονόματα των data-properties της κλάσης “Product” στην οντολογία.
   - Αν η λεξικογραφική ομοιότητα της τιμής κάποιας “class” με κάποιο data-property είναι μεγαλύτερη ή ίση από 0.5 τότε το κείμενο του “class” attribute μπαίνει σαν τιμή στο συγκεκριμένο data-property. Η τιμή 0.5 επιλέχθηκε γιατί από δοκιμές φάνηκε ότι αν είναι μεγαλύτερη (π.χ.0.6) τότε δεν θα προστίθενται τιμές ιδιοτήτων στην οντολογία, ενώ θα έπρεπε. Από την άλλη, αν το όριο ήταν 0.4 θα εισάγονταν τιμές ιδιοτήτων προϊόντων σε λάθος ιδιότητες. Η τιμή 0.5 φάνηκε ότι έχει καλά αποτελέσματα στις περισσότερες περιπτώσεις. Επίσης, πρέπει να σημειωθεί ότι εκτελείται μόνο λεξικογραφική ομοιότητα. Αυτό γίνεται (1) γιατί θα ήταν πολύ αργό να εκτελεστεί σημασιολογική ομοιότητα (όπως περιγράφτηκε στην παράγραφο 5.3.1: Αλγόριθμοι Ομοιότητας) για κάθε ιδιότητα προϊόντος στην ιστοσελίδα με κάθε data-property της κλάσης “Product” στην οντολογία, (2) γιατί οι ιδιότητες προϊόντων περιγράφονται με λέξεις όπως

"product-desc", "prd-name" κτλ. που δεν αποτελούν έγκυρες λέξεις στο WordNet. Επίσης πρέπει να αναφέρουμε ότι η ιδιότητα “hasCurrency” παίρνει τιμή μόνο αν πάρει τιμή και το χαρακτηριστικό “hasPrice” καθώς στις ιστοσελίδες το νόμισμα δίνεται μαζί με την τιμή. Αν το νόμισμα στην ιστοσελίδα έχει τιμή στο σύνολο {"$","£","euro","pound","EUR","R","GBP", "€", "ZAR"}, τότε παίρνει μια από αυτές, διαφορετικά το “hasCurrency” δεν παίρνει τιμή στην οντολογία. Η λίστα με τα πιθανά νομίσματα μπορεί εύκολα να επεκταθεί, ώστε να περιλαμβάνει και όλα τα υπόλοιπα σύμβολα.

- Διαφορετικά, αν η τιμή κάποιου “class” attribute δεν έχει ομοιότητα με κανένα data-property, τότε το κείμενο μπαίνει ως τιμή στο data-property “hasDescription”. Αυτό γίνεται ώστε όσες τιμές χαρακτηριστικών προϊόντων στην ιστοσελίδα δεν μπουν σε κανένα data-property, να μην χαθούν αλλά να είναι διαθέσιμα στο “hasDescription”. Πρέπει να σημειωθεί ότι επειδή στις html σελίδες τα χαρακτηριστικά ανακτώνται σαν String δεδομένα, πρέπει να γίνει μετατροπή σε float τιμές, όπου τα data-properties είναι τύπου float (πχ. “hasPrice”) ώστε η οντολογία να παραμείνει συνεπείς μετά την εισαγωγή των νέων στιγμιοτύπων προϊόντων.

5. Μετά την δημιουργία ενός νέου προϊόντος στην οντολογία, ελέγχεται ότι δεν υπάρχει στην οντολογία και άλλο προϊόν που να περιέχει πανομοιότυπες τιμές στα data-properties και object-properties. Εφόσον δεν υπάρχουν ids, όπως στις σχεσιακές βάσεις, που να επιβάλλουν την μοναδικότητα των στιγμιοτύπων στην οντολογία, ο μόνος τρόπος να διαπιστωθεί ότι δυο στιγμιότυπα είναι ίδια στην οντολογία (duplicate entries), είναι η σύγκριση των data-properties και object-properties που έχουν, και μόνο αν αυτά είναι ακριβώς ίδια μιλάμε για το ίδιο προϊόν, οπότε το ένα πρέπει να διαγραφεί. Η αναζήτηση για ύπαρξη ίδιου προϊόντος μπορεί να γίνει μόνο όταν έχουν ανακτηθεί όλες οι τιμές ιδιοτήτων προϊόντος από την ιστοσελίδα και αυτό γίνεται στο τέλος όταν πλέον έχει δημιουργηθεί το στιγμιότυπο προϊόντος και μαζί με αυτό και οι ιδιότητές του. Αυτά δεν είναι γνωστά εκ των προτέρων,

αλλά μόνο αφού δημιουργηθεί το στιγμιότυπο. Συνεπώς δεν μπορούμε να αποφύγουμε την δημιουργία του και την μετέπειτα καταστροφή του (αν υπάρχει ήδη στην οντολογία).

6. Αποθήκευση των αλλαγών στην οντολογία.

Παράδειγμα ανάκτησης και αποθήκευσης προϊόντος στην οντολογία

Στο Σχήμα 5.4 δίνεται ένα παράδειγμα (http://www.debenhams.com/home-furniture/christmas-decorations) στοιχείου πίνακα μιας ιστοσελίδας που περιλαμβάνει ένα προϊόν και τα χαρακτηριστικά του. Αρχικά παρατηρούμε ότι το στοιχείο αυτό περιγράφει ένα προϊόν γιατί έχει <td class = “product_detail”>, το οποίο σημαίνει ότι η “Μέθοδος Ανάκτησης Προϊόντων” θα το θεωρήσει προϊόν επειδή το “class” attribute περιλαμβάνει το “product”. Εφόσον το θεωρεί προϊόν στη συνέχεια θα ανακτήσει όλα τα εμφωλευμένα στοιχεία του <td>, που είναι όλα τα υπόλοιπα που εμφανίζονται. Από το κάθε εμφωλευμένο στοιχείο, η “Μέθοδος Αποθήκευσης Προϊόντων” θα εξετάσει την τιμή του attribute “class”, δηλαδή τα “product_summary”, “brand_name”, “product_name” κτλ. Θα συγκριθεί το καθένα από αυτά με τα data-properties της κλάσης “Product”, και αν βρεθεί λεξικογραφική ομοιότητα >=0.5 , θα προστεθεί στην οντολογία. Αυτό αποτελεί ένα σενάριο όπου το σύστημα θα εκτελεστεί επιτυχώς και θα ανακτηθούν τα προϊόντα από την ιστοσελίδα.

```
<td class="product_detail">
<a
href="/webapp/wcs/stores/servlet/prod_10001_10001_268100570720_-
1?breadcrumb=Home%7EHome+%26amp%3B+furniture%7EChristmas+decorati
ons" name="&amp;lid=prod_268100570720-Discount"
onclick="sc_trackLink('&amp;lid=268100570720-Discount');">
<imgsrc="http://debenhams.scene7.com/is/image/Debenhams/268100570
720?$PSPMedium$" alt="" />
<div class="product_summary">
 <div class="brand_name">Debenhams</div>
 <div class="product_name">Pack of twelve gold Christmas
crackers</div>
 <div class="product_price">Was &pound;17.00</div>
  <div class="product_price latest">Now &pound;8.50</div>
    <div class="product_rating">
    <img title="0.0 out of 5.0" alt="0.0 out of 5.0"
src="/images/bv/rating-0_0.gif" />
    <span class="product_reviews">0 reviews</span>
    </div>
</div>
</a>
</td>
```

Σχήμα 5.4 : Παράδειγμα περιγραφής προϊόντος με σημασιολογική πληροφορία σε κελί ενός πίνακα

Οι μέθοδοι που χρησιμοποιούνται για την ανάκτηση από ιστοσελίδες των προϊόντων και των χαρακτηριστικών τους βασίζεται στο γεγονός ότι οι ιστοσελίδες έχουν δημιουργηθεί με τέτοιο τρόπο προγραμματιστικά, ώστε τα στοιχεία και τα attributes να περιέχουν μια περιγραφή που να αποδίδει τη σημασιολογία τους. Για παράδειγμα, στο Σχήμα 5.4 βλέπουμε ότι υπάρχουν “class” attributes που περιγράφουν τα χαρακτηριστικά του προϊόντος και έτσι η το σύστημα μπορεί να το ανακτήσει. Αν μια ιστοσελίδα δεν περιλαμβάνει attributes που να περιγράφουν ένα προϊόν τότε η μεθοδολογία που χρησιμοποιεί το σύστημά μας δεν θα μπορέσει να το ανακτήσει (Σχήμα 5.5).

```
 <td align="center" class="smallText" width="33%" valign="top"
height="100%">
<table  cellspacing="0" cellpadding="2" border="0" height="100%">
  <tr>
   <td class="smallText" align="center">
<table width="100" border="0" cellspacing="0" cellpadding="0"
class="image border">
 <tr>
  <td width="33%" align="center">
  <a href="http://www.wineweb.co.za/french-champagne-follet-ramillon-
brut-tradition-p-1479.html"><img src="images/imagecache/Follet-
Tradition-new-b.jpg" border="0" alt="Follet Ramillon Brut Tradition"
title=" Follet Ramillon Brut Tradition " width="75" height="100"
class="ProductImage"></a>
 </td>
</tr>
</table>
</td>
</tr>
<tr> <td align="center"><a class="ProductLink"
href="http://www.wineweb.co.za/french-champagne-follet-ramillon-brut-
tradition-p-1479.html">Follet Ramillon Brut Tradition</a></td></tr>
<tr>
<td align="center" class="ProductInfo">French Champagne </td></tr>
<tr><td align="center" class="ProductInfo">R282.81</td></tr><tr>
<td align="center">
<a class="ProductLink" href="http://www.wineweb.co.za/french-champagne-
follet-ramillon-brut-tradition-p-1479.html">...more Info</a>
</td>
</tr>
<tr>
<td align="center" height="100%" valign="bottom">
<form name="cart_quantity"
```

Σχήμα 5.5 : Παράδειγμα περιγραφής προϊόντος σε πίνακα με μη σημασιολογική πληροφορία

Όπως βλέπουμε στο Σχήμα 5.5 (http://www.wineweb.co.za/french-champagne-c-96.html) δεν υπάρχει κάποιο <table> ή <td> attribute που να αναδεικνύει ότι αυτός ο πίνακας περιλαμβάνει περιγραφή προϊόντος. Επιπλέον, τα χαρακτηριστικά του προϊόντος όπως η ονομασία και η τιμή, προσδιορίζονται και τα δυο με τιμή class = “ProductInfo”, το οποίο δεν δίνει κάποια σημασιολογική πληροφορία, με αποτέλεσμα να μην μπορεί να ανακτηθεί το προϊόν από το σύστημα.

5.4 Οι Τεχνολογίες Υλοποίησης

Το σύστημα έχει αναπτυχθεί στη γλώσσα προγραμματισμού Java, στο περιβάλλον ανάπτυξης NetBeans 7.0. Συγκεκριμένα, τα εργαλεία που χρησιμοποιήθηκαν είναι τα εξής :

- HtmlCleaner 2.2 : Είναι πρόγραμμα ανοιχτού κώδικα γραμμένο σε Java. Για ένα δοσμένο HTML έγγραφο το HtmlCleaner αναδιοργανώνει επιμέρους στοιχεία (elements) και παράγει ένα καλά δομημένο XML έγγραφο. Είναι ένα εργαλείο που χρησιμοποιείται κατά την σάρωση των ιστοσελίδων, γιατί η HTML που βρίσκεται στον Ιστό είναι συνήθως κακογραμμένη και ακατάλληλη για περαιτέρω επεξεργασία [19].
- OWL API 3.2.4 : Είναι ένα Java API ανοιχτού κώδικα για την δημιουργία, το χειρισμό και το serializing OWL οντολογιών. Η τελευταία έκδοση του API εστιάζεται στην OWL 2 [20]. Στην εφαρμογή μας, το OWL API χρησιμοποιείται προγραμματιστικά στο χειρισμό της οντολογίας “Buyer_Knowledge_Base.owl” , για την φόρτωση της οντολογίας, την διάσχισή της με σκοπό την ανάκτηση συγκεκριμένων κλάσεων και object/data properties, την δημιουργία και αποθήκευση νέων στιγμιοτύπων και κλάσεων.
- WordNet 2.1 : Είναι ένα ελεύθερο και δημόσια διαθέσιμο αγγλικό λεξικό όρων, που περιέχει ουσιαστικά, ρήματα, επίθετα και επιρρήματα ομαδοποιημένα σε σύνολα συνωνύμων, το καθένα να εκφράζει διαφορετική έννοια. Τα σύνολα συνωνύμων είναι διασυνδεδεμένα με σημασιολογικές και λεξικολογικές σχέσεις. Η δομή του WordNet το κάνει χρήσιμο εργαλείο για την υπολογιστική γλωσσολογία και την επεξεργασία φυσικής γλώσσας [21]. Στην εφαρμογή μας, το WordNet χρησιμοποιείται από τους αλγορίθμους, για την εύρεση σημασιολογικής ομοιότητας μεταξύ δυο όρων, της κλάσης κατηγορίας προϊόντος που υπάρχει στην οντολογία, και της κατηγορίας προϊόντος που υπάρχει στην ιστοσελίδα. Όπου δεν μπορεί να εφαρμοστεί η σημασιολογική ομοιότητα επειδή κάποιος όρος εισόδου δεν αποτελεί έγγυρη λέξη, τότε χρησιμοποιούνται οι

αλγόριθμοι λεξικογραφικής ανάλυσης για να βρεθεί η ομοιότητα μεταξύ τους.

Επιπλέον, έχουμε χρησιμοποιήσει το Protégé 3.4.6 για οπτικοποίηση της οντολογίας “Buyer_Knowledge_Base.owl”, ώστε να δούμε τις αλλαγές που πραγματοποιούνται σε αυτή μετά από κάθε εκτέλεση της εφαρμογής.

Τέλος, πρέπει να αναφερθεί ότι οι αλγόριθμοι λεξικογραφικής και σημασιολογικής ομοιότητας, που χρησιμοποιούνται στην εφαρμογή είναι υλοποιημένοι από την εργασία [22].

5.5 Σχετικές Εργασίες

Στην βιβλιογραφία υπάρχει ένα πλήθος εργασιών που ασχολούνται με την μάθηση οντολογιών, οι οποίες διακρίνονται με βάση τον τύπο εισόδου, δηλαδή την πηγή από όπου εξάγεται η πληροφορία για την μάθηση της οντολογίας. Οι τύποι εισόδου μπορεί να είναι αδόμητα, ημι-δομημένα ή δομημένα δεδομένα. Αδόμητα δεδομένα είναι το απλό κείμενο όπως βιβλία και περιοδικά, ημι-δομημένα δεδομένα είναι κείμενο σε HTML, XML αρχεία, ενώ δομημένα είναι οι βάσεις δεδομένων και τα λεξικά. Εδώ θα γίνει μια συνοπτική περιγραφή των εργασιών μάθησης οντολογίας από ημι-δομημένα δεδομένα, γιατί αυτό είναι το πεδίο εφαρμογής που μας ενδιαφέρει. Η δημιουργία οντολογιών από ημι-δομημένα δεδομένα χρησιμοποιεί τεχνικές εξόρυξης δεδομένων (data mining) και εξόρυξης περιεχομένου ιστού (web content mining).

Το [57] και [58] χρησιμοποιούν την δομή των ιστοσελίδων για τη δημιουργία ενός πίνακα βάσης δεδομένων, στη συνέχεια χρησιμοποιούν μέθοδο ομαδοποίησης (clustering method) για την δημιουργία των οντολογιών. Χρησιμοποιούν τη δομή των HTML αρχείων με κάποια γλωσσολογικά χαρακτηριστικά για τον προσδιορισμό υποψήφιων εννοιών.

Η εργασία στο [59], [61] ανέπτυξε ένα σύστημα που μαθαίνει από html σελίδες να δημιουργεί ταξινομία (δηλαδή οντολογία) χρησιμοποιώντας μόνο τη δομή των σελίδων. Το OntoMiner περιλαμβάνει αυτόματες τεχνικές για δημιουργία και

συμπλήρωση οντολογίας σε εξειδικευμένο τομέα, με την οργάνωση και την εξόρυξη ενός συνόλου σχετικών και επικαλυπτόμενων taxonomy-directed ιστοσελίδων, που παρέχονται από τον χρήστη και χαρακτηρίζουν το πεδίο ενδιαφέροντος. Μια taxonomy-directed ιστοσελίδα, είναι μια ιστοσελίδα που περιέχει τουλάχιστον μια ταξινομία για την οργάνωση των εννοιών της και παρουσιάζει τα στιγμιότυπα που ανήκουν σε μια έννοια με συγκεκριμένο οργανωμένο τρόπο (όπως επιστημονικά, ειδήσεις και ταξίδια).

Στο [60] παρουσιάζεται ένα σύστημα δημιουργίας οντολογίας από ένα σύνολο σελίδων Ιστού συγκεκριμένου πεδίου και ενός συνόλου εννοιών-βάσης (seed concepts). Χρησιμοποιούνται δυο συμπληρωματικές προσεγγίσεις. Η πρώτη προσέγγιση χρησιμοποιεί την δομή των φράσεων που εμφανίζονται σε κεφαλίδες HTML εγγράφων, ενώ η δεύτερη χρησιμοποιεί την ιεραρχική δομή των HTML κεφαλίδων για τον εντοπισμό νέων εννοιών και των σχέσεων ταξινομίας που υπάρχουν μεταξύ των seed concepts και μεταξύ τους. Το σύστημα χρησιμοποιήθηκε για την δημιουργία οντολογίας στον γεωργικό τομέα.

Η εργασία στο [18] παρουσιάζει μια μεθοδολογία για αυτόματη δημιουργία οντολογίας από σελίδες στον Ιστό που εξάγονται ως αποτέλεσμα αναζήτησης μιας λέξεις-κλειδί σε μηχανή αναζήτησης. Η οντολογία που προκύπτει αναπαριστά μια ταξινομία κλάσεων και δίνει στον χρήστη μια γενική όψη των εννοιών και τις σημαντικότερες ιστοσελίδες που μπορεί να βρεί στον Ιστό για το πεδίο της συγκεκριμένης λέξης-κλειδί.

Στην βιβλιογραφία δεν εντοπίστηκαν πολλές εργασίες που να ασχολούνται αποκλειστικά με την συμπλήρωση οντολογιών (ontology population) με στιγμιότυπα με βάση HTML πίνακες ή λίστες. Μια σχετική εργασία στο [23] παρουσιάζει μια μέθοδο εξαγωγής του προϊόντος από ιστοσελίδα και εισαγωγής του στην σωστή κλάση σε υπάρχουσα οντολογία, κάνοντας αντιστοίχιση των ιδιοτήτων προϊόντος που βρίσκει στην ιστοσελίδα με τις ιδιότητες των κλάσεων στην οντολογία για να βρει το καλύτερο ταίριασμα και να τοποθετήσει το προϊόν σαν στιγμιότυπο στην σωστή κλάση. Ως παράδειγμα χρησιμοποιούνται σελίδες ηλεκτρονικών καταστημάτων που παρουσιάζουν αναλυτικά κάθε IT προϊόν σε

μια σελίδα με πίνακα, ο οποίος περιλαμβάνει όλα τα χαρακτηριστικά του προϊόντος. Η προσέγγιση αυτή διαφέρει από τη δική μας, αφενός γιατί χρησιμοποιείται άλλη δομή πίνακα σαν είσοδο και αφετέρου διότι επικεντρώνεται στην υλοποίηση του ταιριάσματος του web πίνακα με κάποια κλάση στην οντολογία.

ΚΕΦΑΛΑΙΟ 6

ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΣΥΣΤΗΜΑΤΟΣ

Στο κεφάλαιο αυτό θα γίνει αξιολόγηση του συστήματος, ώστε μέσα από τα σενάρια εκτέλεσης να αποτιμηθεί η επίδοσή του ποιοτικά και ποσοτικά. Έτσι, θα περιγραφούν τα σενάρια με βάση τα οποία αξιολογήθηκε το σύστημα και στη συνέχεια θα περιγραφούν τα αποτελέσματα των πειραματικών δοκιμών και κάποια συμπεράσματα.

6.1 Σενάρια Αξιολόγησης

Τα σενάρια που εκτελέστηκαν, είναι επιλεγμένα ώστε να επιδείξουν όσο γίνεται τα χαρακτηριστικά της εφαρμογής και τις επιδόσεις της αναφορικά με τον χρόνο εκτέλεσης.

Το $1^{ο}$ σενάριο αφορά την εξαγωγή προϊόντων από μια σελίδα Ιστού (http://www.skroutz.gr/c/517/camping.html) που περιέχει προϊόντα κατηγορίας “Camping”.

Τα επόμενα σενάρια, $2^{ο}$, $3^{ο}$ και $4^{ο}$ αφορούν την εξαγωγή και αποθήκευση προϊόντων από 5, 10 και 20 αντίστοιχα διαφορετικές σελίδες ηλεκτρονικών καταστημάτων οι οποίες εμφανίζουν τα προϊόντα τους σε html λίστες. Τα σενάρια αυτά μας βοηθούν να αξιολογήσουμε το σύστημα αναφορικά με τον χρόνο εκτέλεσης και την εγκυρότητα των αποτελεσμάτων, όταν έχουμε σαν είσοδο σελίδες που εμφανίζουν τα προϊόντα σε html λίστες.

Το $5^{ο}$ σενάριο είναι η εκτέλεση της εφαρμογής για την εξαγωγή και αποθήκευση στην οντολογία προϊόντων από 10 σελίδες ηλεκτρονικών καταστημάτων που εμφανίζουν τα προϊόντα τους σε μορφή πίνακα. Το σενάριο αυτό δείχνει την επιδόσεις του συστήματος όταν έχουμε σαν είσοδο σελίδες που εμφανίζουν τα προϊόντα σε πίνακα.

Το 6^{ο} σενάριο είναι η εκτέλεση της εφαρμογής για 30 διαφορετικές σελίδες ηλεκτρονικών καταστημάτων, που εμφανίζουν τα προϊόντα τους είτε σε μορφή html λιστών είτε σε μορφή πίνακα.

Το τελευταίο σενάριο αφορά την εκτέλεση των σεναρίων 1 έως 6 σειριακά, το ένα μετά το άλλο και προσθετικά πάνω στην αρχική οντολογία για να δούμε πώς συμπεριφέρεται η εφαρμογή, εάν η οντολογία στην οποία θα προστεθούν νέα στιγμιότυπα προϊόντων, έχει από πριν πολλά στιγμιότυπα και δεν εκτελείται σε άδεια οντολογία, όπως συμβαίνει στα σενάρια 1 εως 6. Με δεδομένο ότι για κάθε νέο στιγμιότυπο που εισάγεται στην οντολογία, ελέγχεται αν υπάρχει ήδη άλλο στιγμιότυπο με το ίδιο όνομα, είναι σημαντικό να δούμε τι γίνεται όταν η οντολογία έχει πολλά στιγμιότυπα.

6.2 Αποτελέσματα Αξιολόγησης

6.2.2 Ποιοτική αξιολόγηση

Τα αποτελέσματα εκτέλεσης των σεναρίων 1 έως 6 παρουσιάζονται στον Πίνακα 6.1. Σύμφωνα με τα αποτελέσματα αυτά, το σύστημα έχει αρκετά καλή επίδοση όσον αφορά στον εντοπισμό και την εξαγωγή των προϊόντων από σελίδες ηλεκτρονικών καταστημάτων που εμφανίζονται σε μορφή πίνακα ή λίστας. Έτσι, στο 6^{ο} σενάριο, από το οποίο δημιουργούνται 692 νέα στιγμιότυπα, οι 576 είναι πραγματικά προϊόντα και μόνο τα 116 είναι στιγμιότυπα που αναπαριστούν προϊόντα ενώ δεν θα έπρεπε, precision 83%. Το χαμηλότερο precision εμφανίζεται στο 5^{ο} σενάριο (58%) αλλά και το 3^{ο} σενάριο (64%). Αυτό συμβαίνει γιατί οι σελίδες Ιστού είναι έτσι δημιουργημένες, ώστε περιέχουν πολλούς πίνακες και/ή λίστες που παρουσιάζουν καταλόγους κατηγοριών προϊόντων, ή πληροφορίες πλοήγησης, με αποτέλεσμα ο αλγόριθμος εντοπισμού προϊόντων που χρησιμοποιεί το σύστημα να θεωρεί κάποια από αυτά ως πραγματικά προϊόντα. Αναφορικά με το recall που έχει ο αλγόριθμος εντοπισμού προϊόντων, από τα πειράματα που εκτελέστηκαν φάνηκε ότι εντοπίζονται και ανακτώνται όλα τα προϊόντα που εμφανίζονται σε πίνακες ή λίστες (δηλαδή δεν μένει κάτι απ'έξω ενώ δεν θα έπρεπε), αν οι δομές αυτές έχουν δημιουργηθεί σε μορφή που να

μπορεί να επεξεργαστεί το σύστημα, όπως αυτή περιγράφτηκε στο προηγούμενο κεφάλαιο. Αν η μορφή παρουσίασης των προϊόντων στις σελίδες Ιστού είναι σύμφωνη με τις προϋποθέσεις που περιγράφονται από το σύστημα, τότε θα ανακτηθούν όλα τα προϊόντα, διαφορετικά αν η μορφή δεν είναι κατάλληλη πιθανότατα δεν θα ανακτηθεί κανένα ή αν ανακτηθεί δεν θα έχει τιμές στις ιδιότητες των προϊόντων στην οντολογία.

Πίνακας 6.1 : Αποτελέσματα αξιολόγησης στιγμιοτύπων

ΣΥΝΟΛΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΞΙΟΛΟΓΗΣΗΣ ΓΙΑ ΣΤΙΓΜΙΟΤΥΠΑ ΠΡΟΪΟΝΤΩΝ					
ΣΕΝΑΡΙΟ	**# ΙΣΤΟΣΕΛΙΔΩΝ**	**# ΝΕΩΝ ΣΤΙΓΜΙΟΤΥΠΩΝ**	**# ΠΡΑΓΜΑΤΙΚΩΝ ΠΡΟΪΟΝΤΩΝ**	**# ΜΗ ΠΡΑΓΜΑΤΙΚΩΝ ΠΡΟΪΟΝΤΩΝ**	**PRECISION ΣΤΙΓΜΙΟΤΥΠΩΝ**
1ο	1	18	18	0	1
2ο	5	102	96	6	96/102 = 0.94
3ο	10	226	145	81	145/226 = 0.64
4ο	20	428	343	85	343/428 = 0.80
5ο	10	234	138	96	138/234 = 0.58
6ο	30	692	576	116	576/692 = 0.83

Εκτός από την αξιολόγηση του συστήματος αναφορικά με το αν εντοπίζει και εξάγει σωστά τα προϊόντα, από πίνακες και λίστες σελίδων Ιστού, ενδιαφέρον παρουσιάζει και η αξιολόγηση αναφορικά με την σωστή εξαγωγή της κατηγορίας των προϊόντων από σελίδες ηλεκτρονικών καταστημάτων. Επίσης, είναι σημαντικό αν η κατηγορία προϊόντων αποθηκεύεται στην σωστή θέση στην οντολογία προϊόντων που χρησιμοποιείται. Τα αποτελέσματα αυτής της αξιολόγησης παρουσιάζονται στον Πίνακα 6.2.

Πίνακας 6.2 : Αποτελέσματα αξιολόγησης κατηγοριών προϊόντων

ΑΞΙΟΛΟΓΗΣΗ ΓΙΑ ΚΑΤΗΓΟΡΙΕΣ ΠΡΟΪΟΝΤΩΝ			
ΣΕΝΑΡΙΟ	**# ΙΣΤΟΣΕΛΙΔΩΝ**	**# ΣΩΣΤΩΝ ΕΙΣΑΓΩΓΩΝ ΚΑΤΗΓΟΡΙΑΣ ΠΡΟΪΟΝΤΩΝ ΣΤΗΝ ΟΝΤΟΛΟΓΙΑ**	**# ΣΩΣΤΩΝ ΑΝΑΚΤΗΣΕΩΝ ΚΑΤΗΓΟΡΙΑΣ ΑΠΟ ΙΣΤΟΣΕΛΙΔΕΣ**
1ο	1	0	1
2ο	5	1/5 = 0.2	4/5 = 0.8
3ο	10	2/10 =0.2	1
4ο	20	5/20 = 0.25	16/20 = 0.8
5ο	10	4/10 = 0.4	1
6ο	30	12/30 = 0.4	29/30 = 0.96

Από τα αποτελέσματα του Πίνακα 6.2 γίνεται φανερό ότι ενώ οι κατηγορίες προϊόντων ανακτώνται στις περισσότερες περιπτώσεις σωστά από τις ιστοσελίδες, δηλαδή # ΣΩΣΤΩΝ ΑΝΑΚΤΗΣΕΩΝ ΚΑΤΗΓΟΡΙΑΣ ΑΠΟ ΙΣΤΟΣΕΛΙΔΕΣ έχουν τιμή από 0.8 έως 1, εν τούτοις η τοποθέτησή τους στη σωστή θέση στην οντολογία εμφανίζει κάποιο πρόβλημα, διότι # ΣΩΣΤΩΝ ΕΙΣΑΓΩΓΩΝ ΚΑΤΗΓΟΡΙΑΣ ΠΡΟΪΟΝΤΟΣ ΣΤΗΝ ΟΝΤΟΛΟΓΙΑ είναι της τάξης του 0.2 με 0.4. Αυτό σημαίνει ότι η "*Μέθοδος Προσδιορισμού Κατηγορίας Προϊόντων*" (Κεφάλαιο 5.3.2) έχει αρκετά καλά αποτελέσματα. Από την άλλη, η μέθοδος "*Αποθήκευσης Κατηγορίας Προϊόντων*" (Κεφάλαιο 5.3.3) που εισάγει την κατηγορία προϊόντων στην οντολογία με χρήση των "Αλγορίθμων Ομοιότητας" (Κεφάλαιο 5.3.1) δεν έχει ικανοποιητικά αποτελέσματα. Η κατηγορία προϊόντων δεν τοποθετείται σωστά και στην ουσία για κάθε κατηγορία που συναντάται στις ιστοσελίδες δημιουργείται μια νέα κλάση, ως υποκλάση της "ProductCategory". Ο λόγος για τον οποίο συμβαίνει αυτό είναι γιατί στον Ιστό, κάθε ηλεκτρονικό κατάστημα έχει τον δικό του τρόπο να παρουσιάζει τα προϊόντα σε κατηγορίες, με διάφορες ονομασίες που δεν υπόκεινται σε κάποιο μοτίβο. Έτσι, οι αλγόριθμοι

σημασιολογικής και λεξικογραφικής ομοιότητας που χρησιμοποιούνται από το σύστημα δεν εντοπίζουν τη συσχέτιση που μπορεί να έχει μια ονομασία κατηγορίας προϊόντων στην ιστοσελίδα (π.χ. Underwear) με την κατηγορία στην οντολογία (π.χ. Cloth), με αποτέλεσμα να φτιάχνει νέα κατηγορία ως κλάση στην οντολογία. Αποτέλεσμα αυτού είναι η οντολογία να γεμίσει με πολλές κλάσεις, μια για κάθε κατηγορία κάθε ηλεκτρονικού καταστήματος.

Στον Πίνακα 6.3 παρουσιάζονται τα επιμέρους αποτελέσματα για το 3^{ο} σενάριο εκτέλεσης , ώστε να αναλυθεί το σύστημα αναφορικά με τιμές που εισάγονται στις ιδιότητες των προϊόντων στην οντολογία. Στον Πίνακα 6.3 τα Α, Β είναι σύνολο ιδιοτήτων προϊόντων οντολογίας με τις εξής τιμές :

A = { hasCurrency, hasPrice, hasCategory, hasProvider, hasDescription}

B = { hasName, hasPid, hasRelevance, hasTimeValidity, hasOriginCountry, hasDiscount, hasShipment, hasManufacturer}

Από όλα τα πειράματα που εκτελέστηκαν, συμπεριλαμβανομένου και του 3^{ου} που παρουσιάζεται, οι ιδιότητες που εισάγονται στις περισσότερες περιπτώσεις σαν τιμές ιδιοτήτων προϊόντων στην οντολογία, είναι αυτές του συνόλου Α. Αυτό συμβαίνει γιατί οι σελίδες των ηλεκτρονικών καταστημάτων, είναι έτσι φτιαγμένες ώστε να μπορεί να εξαχθεί από λεξιλογική ανάλυση των html στοιχείων των λιστών ή των πινάκων, πληροφορία σχετικά με την τιμή ή το νόμισμα πώλησης προϊόντων. Οι ιδιότητες της κατηγορίας Β γενικά δεν ανακτώνται, εκτός από την “hasName”, που πολλές φορές εισάγεται σωστά ως αποτέλεσμα του λεξιλογίου περιγραφής της ιδιότητας ως “prd-name”, “product-name” που είναι κοντά λεξιλογικά στο “hasName” της οντολογίας. Συνολικά, με accuracy (που υπολογίζεται ως το πλήθος των ιδιοτήτων που έλαβαν σωστές τιμές προς το συνολικό πλήθος των ιδιοτήτων της κλάσης “Product”) της τάξης του 0.38 αποθηκεύονται τιμές στις ιδιότητες των προϊόντων. Αποθηκεύονται σωστά στην οντολογία οι πληροφορίες της τιμής, του νομίσματος, της κατηγορίας, του παρόχου προϊόντος, ενώ ό,τι άλλη πληροφορία δεν εισάγεται σε άλλη ιδιότητα, αποθηκεύεται σαν τιμή στο “hasDescription”. Έτσι, δεν χάνεται καμιά τιμή ιδιότητας του προϊόντος που ανακτάται, ώστε με επεξεργασία της ιδιότητας

“hasDescription” να μπορεί να ανακτηθεί όλη η πληροφορία για δεδομένο προϊόν.

Πίνακας 6.3 : Επιμέρους αποτελέσματα αξιολόγησης 3ου σεναρίου

ΕΠΙΜΕΡΟΥΣ ΑΠΟΤΕΛΕΣΜΑΤΑ 3ου ΣΕΝΑΡΙΟΥ							
URL	**# ΝΕΩΝ ΣΤΙΓΜΙΟΤΥΠΩΝ**	**# ΠΡΑΓΜΑΤΙΚΩΝ ΠΡΟΪΟΝΤΩΝ**	**# ΜΗ ΠΡΑΓΜΑΤΙΚΩΝ ΠΡΟΪΟΝΤΩΝ**	**ΙΔΙΟΤΗΤΕΣ ΠΡΟΪΟΝΤΟΣ ΜΕ ΣΩΣΤΕΣ ΤΙΜΕΣ**	**ΙΔΙΟΤΗΤΕΣ ΠΡΟΪΟΝΤΟΣ ΜΕ ΛΑΘΟΣ ΤΙΜΕΣ**	**ΙΔΙΟΤΗΤΕΣ ΠΡΟΪΟΝΤΟΣ ΜΕ ΚΕΝΕΣ ΤΙΜΕΣ**	**ACCURACY ΙΔΙΟΤΗΤΩΝ**
1	44	20	24	A	hasPid	B - hasPid	5/13 = 0.38
2	38	12	26	A	0	B	5/13 = 0.38
3	19	12	7	A	0	B	5/13 = 0.38
4	11	11	0	A - {hasPrice,hasCurrency}	hasName	B - hasName + {hasPrice, hasCurrency}	3/13 = 0.23
5	33	11	22	A	0	B	5/13 = 0.38
6	9	9	0	A + hasName	0	B - hasName	6/13 = 0.46
7	18	18	0	A	0	B	5/13 = 0.38
8	24	24	0	A	hasName	B - hasName	5/13 = 0.38
9	23	20	3	A	hasTimeValidity, hasOriginCountry,	B -{ hasTimeValidity,hasOriginCountry}	5/13 = 0.38
10	8	8	0	A	0	B	5/13 = 0.38

6.2.3 Αξιολόγηση επιδόσεων

Ο Πίνακας 6.4 παρουσιάζει τα αποτελέσματα του χρόνου εκτέλεσης των σεναρίων 1 έως 6 ξεχωριστά το καθένα σε σχεδόν άδεια (με 11 στιγμιότυπα) οντολογία. Τα σενάρια εκτελέστηκαν σε υπολογιστή με χαρακτηριστικά : Επεξεργαστής Core 2 Duo 2.20 GHz. και Μνήμη RAM 4.00 GB.

Πίνακας 6.4 : Αποτελέσματα αξιολόγησης επίδοσης

ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΞΙΟΛΟΓΗΣΗΣ ΕΠΙΔΟΣΗΣ					
ΣΕΝΑΡΙΟ	**# ΙΣΤΟΣΕΛΙΔΩΝ**	**# ΝΕΩΝ ΣΤΙΓΜΙΟΤΥΠΩΝ**	**# ΠΡΑΓΜΑΤΙΚΩΝ ΠΡΟΪΟΝΤΩΝ**	**# ΜΗ ΠΡΑΓΜΑΤΙΚΩΝ ΠΡΟΪΟΝΤΩΝ**	**ΧΡΟΝΟΣ ΕΚΤΕΛΕΣΗΣ**
1ο	1	18	18	0	22 ' 33''
2ο	5	102	96	6	1 ' 4''
3ο	10	226	145	81	8' 42''
4ο	20	428	343	85	5' 6''
5ο	10	234	138	96	4' 40''
6ο	30	692	576	116	28' 42''

Από τα αποτελέσματα του Πίνακα 6.4 το πρώτο ενδιαφέρον στοιχείο αποτελεί ο χρόνος εκτέλεσης του 1^{ου} σεναρίου που αποτελείται από μια ιστοσελίδα με 18 προϊόντα. Ο χρόνος εκτέλεσής του υπερβαίνει το χρόνο εκτέλεσης ακόμη και του 4^{ου} σεναρίου που αποτελείται από 20 σελίδες, ενώ είναι κοντά και στον χρόνο εκτέλεσης του 6^{ου} σεναρίου, που αποτελείται από 30 ιστοσελίδες. Αυτό συμβαίνει γιατί η κατηγορία προϊόντων (“Camping”) που εξάγεται από την ιστοσελίδα αποτελεί έγκυρος όρος στο WordNet, με αποτέλεσμα να εκτελούνται οι αλγόριθμοι σημασιολογικής ομοιότητας, οι οποίοι είναι αργοί στην εκτέλεση λόγω του γεγονότος ότι ανατρέχουν σε ολόκληρη την ιεραρχία του WordNet δένδρου. Στα υπόλοιπα σενάρια οι κατηγορίες προϊόντων δεν εντοπίζονται στο WordNet,

επομένως δεν εκτελούνται οι αλγόριθμοι σημασιολογικής ομοιότητας, είτε γιατί αποτελούνται από περισσότερες λέξεις (π.χ. είναι της μορφής “Music_Instruments”) είτε γιατί οι κατηγορίες εκφράζονται στον πληθυντικό αριθμό (π.χ. “Mice” και όχι “Mouse”) και η μέθοδος των αλγορίθμων σημασιολογικής ομοιότητας δεν τις βρίσκει στο WordNet. Ως αποτέλεσμα, γενικά στις περισσότερες περιπτώσεις καταλήγει να εκτελούνται μόνο οι αλγόριθμοι λεξικογραφικής ομοιότητας. Αυτό είναι συνέπεια της πρακτικής παρουσίασης των κατηγοριών προϊόντων των ηλεκτρονικών καταστημάτων, γιατί παρουσιάζουν τις κατηγορίες προϊόντων με λέξεις ή συνδυασμό λέξεων που δεν υπάρχουν στο WordNet, με αποτέλεσμα το σύστημα να εκτελεί μόνο τους αλγορίθμους λεξιλογικής ομοιότητας (Κεφάλαιο 5.3.1) οι οποίοι είναι πολύ γρήγοροι. Γενικά, ο χρόνος εκτέλεσης της εφαρμογής είναι ανάλογος του πλήθους των προϊόντων που θα εξάγει, δηλαδή του πλήθους των ιστοσελίδων που έχει ως είσοδο, αλλά και της δομής κάθε ιστοσελίδας. Ιστοσελίδες με πολλά εμφωλευμένα στοιχεία και πολλά στοιχεία <li>, <td>, <table> θα καθυστερήσουν περισσότερο την εκτέλεση συγκριτικά με άλλες που έχουν πιο “επίπεδη” δόμηση με λιγότερα εμφωλευμένα στοιχεία html λιστών και πινάκων.

Τέλος, στον Πίνακα 6.5 παρουσιάζονται τα αποτελέσματα επίδοσης του 7^{ου} σεναρίου εκτέλεσης. Το σενάριο αυτό περιλαμβάνει την σειριακή εκτέλεση των σεναρίων 1 έως 6, ώστε να δούμε πώς αλλάζει ο χρόνος εκτέλεσης εάν προστίθενται συνεχώς νέα στιγμιότυπα στην οντολογία προϊόντων. Μας ενδιαφέρει να μπορούμε να χρησιμοποιούμε το σύστημα ως εργαλείο για τον διαρκή εμπλουτισμό μιας οντολογίας προϊόντων με νέα στιγμιότυπα. Για να συμβεί αυτό πρέπει να ελέγξουμε ότι η ύπαρξη πολλών στιγμιοτύπων δεν επιφέρει όλο και μεγαλύτερη καθυστέρηση σε κάθε επόμενη εκτέλεση του συστήματος, ως συνέπεια του γεγονότος ότι για κάθε νέο στιγμιότυπο γίνεται σύγκριση με όλα τα υπάρχοντα στην οντολογία στιγμιότυπα, ώστε να μην αποθηκευτούν δυο διαφορετικά προϊόντα ως ένα μοναδικό στιγμιότυπο. Και αν αυτό συμβαίνει, πρέπει να γνωρίζουμε σε ποιο βαθμό αυξάνεται ο χρόνος εκτέλεσης σε κάθε επόμενη προσθήκη νέων προϊόντων. Συμπεραίνουμε ότι ο χρόνος εκτέλεσης αυξάνεται σε μικρό βαθμό αν συγκριθεί με τους χρόνους

εκτέλεσης των σεναρίων 1 εως 6 (Πίνακας 6.4) που εκτελούνταν σε άδεια οντολογία. Για παράδειγμα, το 6ο σενάριο είχε χρόνο εκτέλεσης (28' 42") σε άδεια οντολογία (Πίνακας 6.4) και όταν εκτελέστηκε σε οντολογία με 1014 στιγμιότυπα, εμφάνισε χρόνο εκτέλεσης (37' 55") (Πίνακας 6.5). Μετά την εκτέλεση του 7ου σεναρίου θα υπάρχουν στην οντολογία συνολικά 1695 στιγμιότυπα προϊόντων.

Πίνακας 6.5 : Αποτελέσματα αξιολόγησης επίδοσης 7^{ου} σεναρίου

ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΞΙΟΛΟΓΗΣΗΣ ΕΠΙΔΟΣΗΣ				
ΣΕΝΑΡΙΟ	**ΥΠΟΣΕΝΑΡΙΟ**	**# ΙΣΤΟΣΕΛΙΔΩΝ**	**# ΝΕΩΝ ΣΤΙΓΜΙΟΤΥΠΩΝ**	**ΧΡΟΝΟΣ ΕΚΤΕΛΕΣΗΣ**
7ο	**1ο**	1	18	20 ' 33"
	2ο	5	102	1' 17"
	3ο	10	226	9' 40"
	4ο	20	423	6' 51"
	5ο	10	234	7' 33"
	6ο	30	692	37' 55"

ΚΕΦΑΛΑΙΟ 7

ΣΥΜΠΕΡΑΣΜΑΤΑ

7.1 Τελικά Συμπεράσματα

Σε αυτή την εργασία παρουσιάστηκε μια προσέγγιση για την αυτόματη συμπλήρωση οντολογίας προϊόντων με στιγμιότυπα που εξάγονται από HTML πίνακες και λίστες ηλεκτρονικών καταστημάτων. Οι σελίδες των ηλεκτρονικών καταστημάτων είναι μια πλούσια πηγή πληροφοριών για προϊόντα, και περιλαμβάνουν χαρακτηριστικά όπως είναι η εμπορική ονομασία, οι τιμές και άλλες ιδιότητες των προϊόντων τους. Η πληροφορία αυτή αποτελεί στην πραγματικότητα μια όψη των δεδομένων προϊόντων που είναι αποθηκευμένα στις σχεσιακές βάσεις δεδομένων των ηλεκτρονικών καταστημάτων. Δεδομένου ότι δεν υπάρχει δημόσια πρόσβαση στις σχεσιακές βάσεις των ηλεκτρονικών καταστημάτων, η εκμαίευση όλης της χρήσιμης πληροφορίας για τα προϊόντα μπορεί να γίνει μέσω των HTML σελίδων στις οποίες παρουσιάζονται. Σε αυτό το πλαίσιο, η προτεινόμενη προσέγγιση είναι να εκμεταλλευτεί τη δομή παρουσίασης των προϊόντων σε HTML λίστες και πίνακες σελίδων Ιστού, ώστε να εμπλουτίσει με νέα στιγμιότυπα προϊόντων και να επεκτείνει με νέες κλάσεις κατηγορίας προϊόντων μια υπάρχουσα οντολογία προϊόντων. Η χρήση των δομών HTML λιστών και πινάκων σαν στοιχείο εισόδου είναι κατάλληλη, γιατί περιλαμβάνουν μια δομημένη παρουσίαση των προϊόντων και επιπλέον αποτελεί σχεδιαστική επιλογή στις περισσότερες σελίδες ηλεκτρονικών καταστημάτων για την παρουσίαση των προϊόντων τους σήμερα.

Η μέθοδος ανάκτησης προϊόντων που χρησιμοποιήθηκε βασίζεται σε λεξικογραφική ανάλυση των HTML στοιχείων (tags) που παριστάνουν λίστες και πίνακες και σε λεξικογραφική ανάλυση των τιμών των attributes που εμπεριέχονται σε αυτά, ώστε να ανακτηθούν οι ιδιότητες των προϊόντων. Τα βήματα της υλοποίησης περιλαμβάνουν την σάρωση μιας HTML σελίδας, τον

εντοπισμό της κατηγορίας των προϊόντων που παρουσιάζονται σε αυτή με βάση στοιχεία headers και title, την αποθήκευση της κατηγορίας ως κλάση στην οντολογία ή τον εντοπισμό της στην ιεραρχία κλάσεων της οντολογίας αν υπάρχει ήδη, την ανάκτηση των προϊόντων και ιδιοτήτων τους με βάση λεξικογραφική ανάλυση των HTML στοιχείων λιστών και πινάκων της σελίδας και τέλος, την αποθήκευση αυτών στην σωστή θέση στην οντολογία.

Μπορεί να διαπιστωθεί, από τα αποτελέσματα αξιολόγησης, ότι η προτεινόμενη προσέγγιση είναι πρακτική και χρήσιμη για την μείωση του χρόνου που καταναλώνεται για την δημιουργία στιγμιοτύπων προϊόντων. Αποφεύγεται η χειροκίνητη εισαγωγή στιγμιοτύπων προϊόντων σε οντολογία, εισάγονται στιγμιότυπα για προϊόντα από κατανεμημένες πηγές σε ένα σημείο συγκέντρωσης. Έτσι, η πληροφορία που είναι αποθηκευμένη στις σχεσιακές βάσεις και γίνεται προσβάσιμη μέσω HTML λιστών και πινάκων των σελίδων Ιστού είναι πλέον διαθέσιμη στον Ιστό και επεξεργάσιμη από μηχανές. Αυτό σημαίνει ότι η εμπλουτισμένη οντολογία προϊόντων μπορεί να χρησιμοποιηθεί από λογισμικούς πράκτορες (π.χ. shopbots) ή άλλες οντότητες μιας εικονικής αγοράς, ώστε να εκτελούν σύνθετες εργασίες και επερωτήσεις στην οντολογία προϊόντων, με σκοπό την παροχή καλύτερης πληροφόρησης στους χρήστες για τα προϊόντα και τα χαρακτηριστικά τους.

7.2 Μελλοντικές Κατευθύνσεις

Μελλοντικές κατευθύνσεις έρευνας στα πλαίσια της προσέγγισης που αναπτύχθηκε σε αυτή την εργασία αποτελούν οι παρακάτω:

- Η εύρεση πιο αποτελεσματικής μεθόδου αποθήκευσης κατηγορίας προϊόντων στην οντολογία. Ο αλγόριθμος που χρησιμοποιείται στην εργασία για την εύρεση της σωστής θέσης μιας κατηγορίας προϊόντων που εξάγεται από ιστοσελίδα στην ιεραρχία κλάσεων της οντολογίας προϊόντων, δεν έχει πολύ καλά αποτελέσματα και δημιουργεί νέα κλάση για κάθε κατηγορία χωρίς να εντοπίζει πάντα ότι η κατηγορία προϊόντων υπάρχει ήδη στην οντολογία. Πρέπει να βελτιωθεί ο τρόπος χρήσης των αλγορίθμων ομοιότητας, ώστε να

βρεθεί πιο αποτελεσματικός τρόπος αντιστοίχισης της κατηγορίας προϊόντος που ανακτάται από την ιστοσελίδα με τη σωστή θέση τοποθέτησής της στην οντολογία.

- Τα ονόματα των νέων στιγμιοτύπων δημιουργούνται με την χρήση μιας Random γεννήτριας τυχαίων αριθμών τύπου long, ώστε να αποφευχθεί η αναπαράσταση δυο διαφορετικών προϊόντων με το ίδιο στιγμιότυπο στην οντολογία. Ενδεχομένως, να υπάρχει άλλη καλύτερη μέθοδος δημιουργίας παγκόσμια μοναδικών αναγνωριστικών προϊόντων για την δημιουργία στιγμιοτύπων.
- Θα ήταν χρήσιμο αν μπορούσε να συνδυαστεί η δημιουργία νέων κλάσεων ως κατηγορίες προϊόντων στην οντολογία, με ταυτόχρονη δημιουργία object-properties ή data-properies κατάλληλων για την κλάση αυτή, ώστε να προκύψει πιο εκφραστική οντολογία. Όμως, τώρα αυτό δεν είναι εφικτό γιατί δεν μπορεί να εξαχθεί τέτοια σύνθετη πληροφορία από ανάλυση των html headers και titles της κάθε ιστοσελίδας.
- Από όλες τις ιδιότητες που περιγράφουν ένα προϊόν στην οντολογία, μόνο ένα μικρό μέρος, που σχετίζονται με πληροφορίες τιμής, νομίσματος, παρόχου και κατηγορίας προϊόντος, παίρνουν πραγματικά τιμές κατά την αυτόματη εισαγωγή νέων προϊόντων από ιστοσελίδες. Οι υπόλοιπες παραμένουν κενές, είτε γιατί δεν παρουσιάζονται στις HTML λίστες και πίνακες των ιστοσελίδων, είτε γιατί ο τρόπος παρουσίασης των HTML attributes που αναπαριστούν τις ιδιότητες προϊόντων στην ιστοσελία δεν είναι σε μορφή που μπορεί να εντοπίσει το σύστημα, με την χρήση μόνο της λεξικογραφικής ομοιότητας που εκτελεί. Βέβαια, τα html attributes που δεν αντιστοιχίζονται σε κάποια ιδιότητα προϊόντος στην οντολογία αποθηκεύονται σαν τιμές της ιδιότητας “hasDescrption”, ώστε να μπορούν να επεξεργαστούν και να μην χάνονται. Μια πιθανή βελτίωση θα ήταν να βρεθεί μέθοδος που να αντιστιχίζει περισσότερα html attributes σε ιδιότητες προϊόντων στην οντολογία από ότι συμβαίνει τώρα.

- Τέλος, δεν υπάρχει υλοποιημένη (built-in) μέθοδος στο OWL API 3.2.4 για την αποθήκευση τιμών τύπου date στις ιδιότητες των προϊόντων στην οντολογία (π.χ. hasValidityTime). Έτσι, το hasValidityTime όταν παίρνει τιμές ως αποτέλεσμα της εκτέλεσης του συστήματος αυτές είναι τύπου String ενώ θα έπρεπε να εισάγονται τιμές τύπου date. Απαιτείται ίσως μια custom υλοποίηση που να δημιουργεί αντικείμενο τύπου date, ώστε να είναι εφικτή η αποθήκευση τιμών ιδιοτήτων τύπου date στην οντολογία.

ΟΡΟΛΟΓΙΑ

Ξένος Όρος	Ελληνικός Όρος
Semantic Web	Σημασιολογικός Ιστός
World Wide Web	Παγκόσμιος Ιστός
Ontology Learning	Μάθηση Οντολογίας
Electronic Marketplace	Εικονική Αγορά
B2B	Επιχείρηση-προς-Επιχείρηση
Instance	Στιγμιότυπο
Concept	Έννοια
Reasoning	Συλλογιστική
Ontology Population	Συμπλήρωση Οντολογίας

ΑΝΑΦΟΡΕΣ

1. T. Berners-Lee, J. Hendler, O. Lassila, “The Semantic Web”, Scientific American, May 2001.
2. W3C Semantic Web Activity, available at http://www.w3.org/2001/sw/, retrieved 28th November 2011
3. W3C, ‘Resource Description Framework (RDF)’, available at http://www.w3.org/RDF/, retrieved 28th November 2011.
4. Web Ontology Language (OWL), available at http://www.w3.org/2004/OWL/
5. Troy J., Strader, and M. J. Shaw, ‘Electronic Markets: Impact and Implications’, Handbook on Electronic Commerce, Springer-Verlag, 1999.
6. Martin Grieger, “Electronic marketplaces: A literature review and a call for supply chain management research”, European Journal of Operational Research 144 (2003) 280–294, www.elsevier.com/locate/dsw
7. Efraim Turban, et al. “Electronic Commerce” Chapter 2, Pearson Prentice Hall, 2008.
8. Chiu Dickson K.W., Poon, Joe Kit Man, Lam Wai Chun, Tse Chi Yung, Sui William Hi Tai, Poon Wing Sze, “HOW ONTOLOGIES CAN HELP IN AN E-MARKETPLACE”, ECIS 2005 Proceedings (2005)
9. Hepp, M., Leukel, J., Schmitz, V. A. Quantitative Analysis of eCl@ss, UNSPSC, eOTD, and RNTD: Content, Coverage, and Maintenance, 2007.
10. Naming and Addressing: URIs, URLs…, available at http://www.w3.org/Addressing/
11. Extensible Markup Language (XML), available at www.w3.org/XML/
12. Ícaro Medeiros, “Ontology Learning”, CIn – UFPE, September 30, 2008
13. United Nations Development Programme: United Nations Standard Products and Services Code (UNSPSC), available at http://www.unspsc.org/
14. eCl@ss , available at http://www.eclass.de/

15. Hepp, Martin F., Product Representation in the Semantic Web (April 28, 2004).

16. Gomez-Perez, A., Manzano-Macho, D. 2003. OntoWeb Deliverable 1.5: A Survey of Ontology Learning Methods and Techniques. Universidad Politecnica de Madrid. Asunción Gómez-Pérez, David Manzano-Macho, “OntoWeb : A survey of ontology learning methods and techniques”

17. Maedche, A. and Staab, S. 2001. Ontology Learning for the Semantic Web. In IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2). Alexander Maedche, Steffen Staab “Learning Ontologies for the Semantic Web”

18. Sanchez, D., and Moreno, A. 2004. Creating ontologies from Web documents. In Recent Advances in Artificial Intelligence Research and Development. IOS Press, Vol.113, pp.11-18. David SÁNCHEZ, Antonio MORENO, “Creating Ontologies from Web documents”

19. HtmlCleaner , available at http://htmlcleaner.sourceforge.net/

20. OWL API, available at http://owlapi.sourceforge.net/index.html

21. WordNet, available at http://wordnet.princeton.edu/

22. Κωνσταντίνος Μ. Κολομβάτσος, “Συγκριτική αξιολόγηση Μεθόδων Λεξικογραφικής και Σημασιολογικής Ομοιότητας”, 2005.

23. Seong-Bae Park, Sang-Soo Kim, Sewook Oh, Zooyl Zeong, Hojin Lee, and Seong Rae Park, “Target Concept Selection by Property Overlap in Ontology Population”

24. Βλαχάβας Ι., Κεφαλάς Π., Βασιλειάδης Ν., Ρεφανίδης Ι., Κόκκορας Φ., Σακελλαρίου Η. “ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ”, 2002, p.287

25. Thomas Gruber, “Toward Principles for the Design of Ontologies Used for Knowledge Sharing”, August, 1993

26. RNTD Development (Rosettanet Technical Dictionary) , available at http://www.rosettanet.org/dnn_rose/Standards/RosettaNetPrograms/Foundati

onalPrograms/CompletedFoundationalPrograms/RNTDDevelopment/tabid/439/Default.aspx

27. eOTD , available at http://www.eccma.org/
28. L. Gravano, P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan, L. Pietarinen, D. Srivastava, “Using q-grams in a DBMS for Approximate String Processing”, IEEE Data Engineering Bulletin, vol. 24, No 4, pp 28-34, 2001.
29. D. Lin, “An Information-Theoretic Definition of Similarity”, In Proceedings of the 15th International Conf. on Machine Learning, pp 296–304. Morgan Kaufmann, San Francisco, CA, 1998.
30. M. A. Jaro, “Advances in record linking methodology as applied to the 1985 census of Tampa Florida”, Journal of the American Statistical Society, vol. 64, pp 1183-1210, 1989.
31. D. D. Palmer, “A Trainable Rule-based Algorithm for Word Segmentation”, Proceedings of the 35th annual meeting on Association for Computational Linguistics, pp 321-328, Madrid, Spain, 1997.
32. RDF Schema, available at http://www.w3.org/TR/rdf-schema/
33. SPARQL, available at http://www.w3.org/TR/rdf-sparql-query/
34. RIF, available at http://www.w3.org/2005/rules/wiki/RIF_Working_Group
35. Van Heck, E., Ribbers, P.M.A., 1997. Economic effects of EM: An analysis of four cases in the dutch flower and transport industries. In: Nunamaker, J.F., Sprague, R.H. (Eds.), Information Systems Organisational Systems and Technology. Proceedings of the 29th Annual Hawaii International Conference on System Sciences. IEEE Computer Society Press, Los Alamitos, pp. 407–415.
36. Martin Hepp, “THE TRUE COMPLEXITY OF PRODUCT REPRESENTATION IN THE SEMANTIC WEB”

37. Obrst, L., Wray, R.E. and Liu, H., Ontological Engineering for B2B E-Commerce. in International Conference on Formal Ontology in Information Systems (FOIS'01), (Ogunquit, Maine, USA, 2001), ACM Press, 117-126.

38. Kashyap, V. (1999).Design and creation of ontologies for environmental information retrieval. In Proceeding of the 12th Workshop on Knowledge Acquisition, Modelling and Management (KAW), Banff, Alberta, Canada.pp. 3–21.

39. Stojanovic, L., N. Stojanovic and R. Volz (2002). Migrating data-intensive web sites into the semantic web. In Proceeding of the 17th ACM Symposium on Applied Computing (SAC'2002), Madrid, Spain. pp. 1100–1107.

40. Rubin, D.L., M. Hewett, D.E. Oliver, T.E Klein and R.B. Altman (2002). Automatic data acquisition into ontologies from pharmacogenetics relational data sources using declarative object definitions and XML. In R.B. Lihue et al. (Eds.), Proceedings of the Pacific Symposium on Biology. pp. 22–34.

41. Astrova, I. (2004). Reverse engineering of relational databases to ontologies. In Proceeding of the 1st European Semantic Web Symposium (ESWS), Heraklion, Greece. LNCS, vol. 3053. pp. 327–341.

42. Tijerino, Y.A, D.W. Embly, D.W. Lonsdale, Y. Ding and G. Nagy (2005). Towards ontology generation from tables. 8(3), 261–285.

43. Astrova, I., and B. Stantic (2005). An HTML forms driven approach to reverse engineering of relational databases to ontologies. In Proceeding of the 23rd IASTED International Conference on Databases and Applications (DBA), Innsbruck, Austria. pp. 246–251.

44. Benslimane, S. M., Malki, M., Rahmouni, M. K., and Benslimane, D. 2007. Extracting personalised ontology from data-intensive web application: An HTML forms-based reverse engineering approach. Informatica, 18, 4 (Dec.2007), 511-534.

45. Schulten, E., H. Akkermans, G. Botquin, M. Dörr, N. Guarino, N. Lopes and N. Sadeh (2001). The E-Commerce Product Classification Challenge. IEEE Intelligent Systems 16 (4), 86-89.

46. K. Bontcheva and H. Cunningham. 2003. The Semantic Web: A New Opportunity and Challenge for HLT. In Proceedings of the Workshop HLT for the Semantic Web and Web Services at ISWC 2003.

47. P. Resnik, “Using Information Content to Evaluate Semantic Similarity in a Taxonomy”, In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, 1995.

48. C. Leacock, M. Chodorow, “Combining local context and WordNet similarity for word sense identification”, In C. Fellbaum, editor, WordNet: An Electronic Lexical Database, pp 265–283, MIT Press, 1998.

49. J. J. Jiang, D. W. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy“, In Proceedings of International Conference Research on Computational Linguistics (ROCLING X), Taiwan, 1997.

50. D. Lin, “An Information-Theoretic Definition of Similarity”, In Proceedings of the 15th International Conf. on Machine Learning, pp 296–304. Morgan Kaufmann, San Francisco, CA, 1998.

51. Z. Wu, M. Palmer, “Verb semantics and lexical selection”, In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp 133–138, Las Cruces, New Mexico, 1994.

52. A. Tversky, “Features of similarity”, Psychological Review, vol. 4, pp 327–352, 1977.

53. R. Rada, H. Mili, E. Bicknell, M. Blettner, ”Development and Application of a Metric on Semantic Nets”, IEEE Transactions on Systems, Man, and Cybernetics, vol. 19, No 1, pp 17–30, 1989.

54. Y. Li, Z. A. Bandar, D. McLean, “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources”, IEEE

Transactions on Knowledge and Data Engineering, vol . 15, No. 4, JULY/AUGUST 2003.

55. E. Agirre, G. Rigau, “A Proposal for Word Sense Disambiguation using Conceptual Distance”, In Proceedings of the First International Conference on Recent Advances in NLP. – Tzigov Chark, Bulgaria, September 1995.

56. E. Agirre, G. Rigau, “Word Sense Disambiguation using Conceptual Density”, In Proceedings of the 16th International Conference on Computational Linguistics, pp 16–22, Copenhagen, 1996.

57. Karoui, L., Aufaure, M., and Bennacer, N. 2004. Ontology Discovery from Web Pages: Application to Tourism. In ECML/PKDD 2004: Knowledge Discovery and Ontologies KDO-2004.

58. Bennacer, N., and Karoui L. 2005. A framework for retrieving conceptual knowledge from Web pages. In Semantic Web Applications and Perspectives, Proceedings of the 2nd Italian Semantic Web Workshop, University of Trento, Trento, Italy.

59. Davulcu, H., Vadrevu, S., and Nagarajan, S. 2004. OntoMiner: Bootstrapping Ontologies From Overlapping Domain Specific Web Sites. In: Poster presentation at the 13th International World Wide Web Conference May 17-22 2004, New York, NY.

60. Hazman, M., El-Beltagy, S. R., and Rafea, A. 2009. Ontology Learning from Domain Specific Web Documents. In International Journal of Metadata, Semantics and Ontologies, Vol. 4, No. 1-2, pp: 24 – 33.

61. Davulcu, H., Vadrevu, S., Nagarajan, S., and Ramakrishnan, I. 2003. OntoMiner: Bootstrapping and Populating Ontologies from Domain Specific Web Sites. In IEEE Intelligent Systems, Vol. 18, No. 5, pp. 24-33.