ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΑΡΧΕΙΟΝΟΜΙΑΣ – ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Εξατομικευμένη ανάκτηση ειδήσεων με τη βοήθεια οντολογιών και εξαγωγής πληροφορίας»

ΤΣΙΦΛΙΔΟΥ ΕΥΘΥΜΙΑ
Α.Μ.:Β00070

Επιβλέπων καθηγητής
Χρήστος Παπαθεοδώρου

Αθήνα, Σεμπτέμβριος 2005

Περιεχόμενα

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κ.Χρήστο Παπαθεοδώρου, για την ουσιαστική συμπαράστασή του σε όλη τη διάρκεια εκπόνησης της πτυχιακής μου εργασίας. Θα ήθελα ακόμη να εκφράσω τις ευχαριστίες μου στον κ. Γιώργο Παλιούρα, διδάκτορα και ερευνητή του Ινστιτούτου Πληροφορικής και Τηλεπικοινωνιών στο Εθνικό Κέντρο Έρευνας Φυσικών Επιστημών «Δημόκριτος», η παρουσία και καθοδήγηση του οποίου υπήρξε καταλυτική προκειμένου να ολοκληρώθεί η προσπάθεια αυτή. Τον ευχαριστώ ιδιαιτέρως.

Θερμές ευχαριστίες θα ήθελα να δώσω και στον επίκουρο καθηγητή του Ιονίου πανεπιστημίου κ. Μανόλη Γεργατσούλη για την συμμετοχή του στην εξέταση και διόρθωση της εργασίας αυτής.

Επίσης, θα ήθελα να ευχαριστήσω τον φοιτητή του τμήματος Πληροφορικής και Τηλεπικοινωνιών του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών Άγγελο Αλεξόπουλο για όλη τη βοήθεια που μου προσέφερε. Πολύτιμη υπήρξε και η παρουσία των φοιτητών Αλέξανδρου Μουζακίδη και Χρήστου Ντούτση του Τμήματος Πληροφορικής του ΤΕΙ Αθηνών. Τέλος, θα ήθελα να ευχαριστήσω τους γονείς και τους φίλους μου για την συμπαράσταση και ειλικρινή κατανόησή τους καθ’ όλη την διάρκεια της εκπόνησης της πτυχιακής εργασίας. Τους ευχαριστώ όλους θερμά.

Περίληψη

Η παρούσα πτυχιακή εργασία πραγματοποιήθηκε υπό την επίβλεψη του εργαστηρίου Τεχνολογίας και Λογισμικού του Ε.Κ.Ε.Φ.Ε. Δημόκριτος με σκοπό την βελτίωση του Συστήματος Εξατομικευμένης Ενημέρωσης (Personalized News Delivery Service), το οποίο αναπτύχθηκε στα πλαίσια έργου του εργαστηρίου, και ως πτυχιακή εργασία του φοιτητή Αλεξόπουλου Άγγελου του Τμήματος Πληροφορικής και Τηλεπικοινωνιών του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών. Το σύστημα προσφέρει υπηρεσίες εξατομικευμένης ενημέρωσης εξάγωντας ειδήσεις από ειδησεογραφικές πηγές στον Παγκόσμιο Ιστό (World Wide Web). Στην παρούσα εργασία περιέχεται μία συνοπτική περιγραφή του συστήματος, το οποίο αποτελείται από δύο εξυπηρετητές, τον Content Server, ο οποίος αναπτύχθηκε από τον φοιτητή Αλεξόπουλο Άγγελο και από τον Personalization Server ο οποίος αναπτύχθηκε σε ανεξάρτητο πρόγραμμα του Ε.Κ.Ε.Φ.Ε «Δημόκριτος». Ακόμα, γίνεται μία σύντομη αναφορά στις υπηρεσίες εξατομικευμένης ενημέρωσης και γενικότερα παρουσιάζονται συστήματα που παρέχουν τις υπηρεσίες αυτές.

Επίσης, πραγματοποιήθηκε μία μελέτη σχετικά με την ανάπτυξη των οντολογιών και την σημασία τους στην εξέλιξη του Σημασιολογικού Ιστού (Semantic Web). Αφού μελετήθηκαν οι σημαντικότερες γλώσσες αναπαράστασης των οντολογιών, στη συνέχεια έγινε μία έρευνα για τα πρότυπα ανάπτυξης οντολογιών και για τις οντολογίες που οφορούν τον ειδησεογραφικό τομέα (News Domain), προκειμένου να χρησιμοποιηθούν για την κατηγοριοποίηση των ειδήσεων στο Σύστημα.

Τέλος, στα πλαίσια της εργασίας μελετήθηκαν διάφορες ειδησεογραφικές πηγές, προκειμένου να συμπεριληφθούν στο Σύστημα Εξατομικευμένης Ενημέρωσης. Για την εξαγωγή των ειδήσεων από τις πηγές αυτές επεκτάθηκαν τα προγράμματα wrappers, τα οποία χρησιμοποιεί το συγκεκριμένο σύστημα για τη λειτουργία αυτή.

1.ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο της πτυχιακής εργασίας

Η παρούσα πτυχιακή εργασία πραγματεύεται την ανάπτυξη οντολογιών, ως εργαλείων τεχνολογικής υποδομής του Σημασιολογικού Ιστού, με σκοπό την διευκόλυνση στην επεξεργασία και την κατηγοριοποίηση ειδησεογραφικού περιεχομένου. Μελετάται η εφαρμογή τους σε Συστήματα Εξατομικευμένης Ενημέρωσης, εφαρμόζοντας μία οντολογία σε ένα αντίστοιχο Σύστημα, το οποίο αναπτύχθηκε στα πλαίσια έργου του τμήματος Τεχνολογιών και Λογισμικού του Ε.Κ.Ε.Φ.Ε «Δημόκριτος».

Στα πλαίσια της εργασίας αυτής πραγματοποιήθηκαν ορισμένες αλλαγές στη Βάση Δεδομένων Περιεχομένου του προαναφερθέντος συστήματος, με αφορμή την προσθήκη επιπλέον ειδησεογραφικών πηγών από τις οποίες εξάγονται οι ειδήσεις του συστήματος. Η εξαγωγή των ειδήσεων πραγματοποιείται με τη βοήθεια των wrappers (προγράμματα εξαγωγής πληροφορίας), πάνω στους οποίους δουλέψαμε προκειμένου να υλοποιηθεί η προσπάθεια της παρούσας πτυχιακής εργασίας.

Στην ενότητα αυτή γίνεται μία εισαγωγή στις έννοιες του Σημασιολογικού Ιστού, της εξατομίκευσης υπηρεσιών (υπηρεσίες στα πλαίσια της ανάπτυξης του Σημασιολογικού Ιστού) και των οντολογιών, μέσων απαραίτητων για την σημασιολογική αναπαράσταση της πληροφορίας.

1.1.1 Προς ένα Σημασιολογικό Ιστό

Μέσα σε λίγο περισσότερο από δέκα χρόνια το Διαδίκτυο κατάφερε να προσελκύσει ένα μεγάλο πλήθος χρηστών, κυρίως χάρη στην ευκολία πρόσβασης και παροχής υπηρεσιών στον Παγκόσμιο Ιστό. Οι χρήστες που προσελκύονται από το Διαδίκτυο πλέον δεν ανήκουν μόνο σε ανθρώπους

της Πανεπιστημιακής κοινότητας ή των εταιρειών πληροφορικής, αντίθετα είναι άνθρωποι διαφόρων ηλικιών και επαγγελμάτων, που βρίσκουν ενδιαφέρον στον Ιστό ο καθένας για εντελώς διαφορετικούς λόγους. Φυσικό επακόλουθο της τόσο μεγάλης προσέλευσης ήταν η παράλληλη αύξηση των ιστοσελίδων που δημοσιεύονται στον Ιστό προκαλώντας αναπόφευκτα την αύξηση της πληροφορίας σε τέτοιο βαθμό ώστε σήμερα να αντιμετωπίζεται το πρόβλημα της *υπερπληροφόρησης* (information overload) στον Παγκόσμιο Ιστό.

Σ' έναν Ιστό, λοιπόν, όπου η πληθώρα πληροφοριών είναι τόσο μεγάλη, οι χρήστες, ιδίως εκείνοι που δεν διαθέτουν τις κατάλληλες γνώσεις και δεν είναι επαρκώς ενημερωμένοι, δυσκολεύονται και πολύ συχνά αποτυγχάνουν να εντοπίσουν τις πληροφορίες που αναζητούν. Παράλληλα, οι απαιτήσεις των χρηστών για καλύτερες μεθόδους αναζήτησης και εύρεσης πληροφορίας διαρκώς μεγαλώνουν αναγκάζοντας τις εταιρείες και τις επιχειρήσεις στο Διαδίκτυο να δώσουν προστιθέμενη αξία στις εφαρμογές τους, αυξάνοντας μ' αυτό τον τρόπο άλλωστε και την καθιερωμένη πελατεία τους.

Επίσης, η σημερινή αναπαράσταση της πληροφορίας στο διαδίκτυο στο μεγαλύτερο μέρος της προορίζεται για κατανόηση και διαχείριση από τους ανθρώπους και όχι από τα υπολογιστικά συστήματα. Τα υπολογιστικά συστήματα δεν μπορούν να επεξεργαστούν έννοιες και σημασιολογίες και κατά συνέπεια να ανταποκριθούν σε σύνθετες αναζητήσεις των χρηστών. Προς την βελτίωση της δόμησης της πληροφορίας, ώστε αυτή να είναι προσπελάσιμη από εφαρμογές υπολογιστών, όχι μόνο για λόγους παρουσίασης της πληροφορίας, αλλά με τελικό στόχο την αυτοματοποίηση, ολοκλήρωση (integration), και επαναχρησιμοποίηση των δεδομένων σε διάφορες εφαρμογές, κατευθύνεται ο Σημασιολογικός Ιστός.

Ο Σημασιολογικός Ιστός (Semantic Web) αποτελεί το επόμενο βήμα του Παγκόσμιου Ιστού. Ένα όραμα και μία πρόταση για την μετεξέλιξη του Διαδικτύου, όπου η πληροφορία αποκτά πλέον δομή και σημασιολογία και ευρύτερος στόχος τίθεται η αποδοτική αναζήτηση και επεξεργασία

δεδομένων. Εμπνευστής του όρου Semantic Web και της υλοποίησης της αρχιτεκτονικής του είναι ο Tim Berners-Lee, σύμφωνα με τον οποίο «ο Σημασιολογικός Ιστός δεν είναι ένας ξεχωριστός ιστός αλλά η επέκταση του συντακτικού ιστού, στον οποίο η πληροφορία είναι καλά καθορισμένη καθιστώντας καλύτερη τη συνεργασία ανθρώπων και υπολογιστών» [Berners-Lee, Lassila, Hendler].

Το όραμα του Σημασιολογικού Ιστού είναι να πραγματοποιείται η επεξεργασία δεδομένων από τα υπολογιστικά συστήματα με τέτοιο τρόπο ώστε να ικανοποιούνται όσο το δυνατόν ακόμα πιο σύνθετες απαιτήσεις των χρηστών.

1.1.2 Εξατομίκευση

Στα πλαίσια της ανάπτυξης ενός Σημασιολογικού Ιστού, όπου το περιεχόμενο, το οποίο διακινείται στο διαδίκτυο, μπορεί να τύχει επεξεργασίας και να κατανοηθεί τόσο από τους ανθρώπους όσο και από τις μηχανές, εντάσσεται και η εμφάνιση των «*τεχνικών εξατομικευμένων υπηρεσιών*». Σε έναν ιστό όπου η πληθώρα πληροφοριών δυσκολεύει την αναζήτηση και την ανάκτηση, η εξατομίκευση (personalization) καλείται να αποτελέσει μία έξυπνη πρόταση, όπου το *περιβάλλον του Παγκόσμιου Ιστού* (*Web environment*) προσαρμόζεται στις προτιμήσεις του χρήστη.

Εξατομίκευση είναι η διαδικασία κατά την οποία συλλέγονται και αποθηκεύονται πληροφορίες αναφορικά με τους χρήστες ενός συστήματος, η ανάλυση αυτής της πληροφορίας και η ανάκτηση των κατάλληλων δεδομένων και πληροφοριών για τον συγκεκριμένο χρήστη την κατάλληλη χρονική στιγμή.

Σήμερα, η δομή του Παγκόσμιου Ιστού αναγκάζει σε μία πλοήγηση από ιστοσελίδα σε ιστοσελίδα προκειμένου να εντοπιστούν οι αναζητούμενες πληροφορίες και ειδήσεις. Μία εξατομικευμένη ιστοσελίδα *φέρνει* τα νέα

και την πληροφορία απευθείας στον χρήστη προσωπικά. Δεν είναι καθόλου περίεργο, λοιπόν, που πολλές επιχειρήσεις στο Διαδίκτυο άρχιζαν να εφαρμόζουν στις ιστοσελίδες τους τεχνικές εξατομικευμένων υπηρεσιών. Ικανοποιώντας τις ανάγκες των πελατών εξασφαλίζεται σε μεγάλο βαθμό η επαναχρησιμοποίηση των υπηρεσιών, ο πελάτης παραμένει περισσότερο στην ιστοσελίδα και κατά συνέπεια αυξάνεται η εμπορική αξία της της ιστοσελίδας.

Πλέον οι χρήστες έχουν συνηθίσει την παρουσία συστημάτων εξατομίκευσης και πολύ συχνά μάλιστα τα θεωρούν δεδομένα. Σύμφωνα με την Global Landscape, το 56% των χρηστών που διαβάζουν ηλεκτρονικές εφημερίδες, σήμερα προτιμούν εξατομικευμένες εφημερίδες. (Αλεξόπουλος, 2003)

1.1.3 Το ζήτημα των οντολογιών στο Σημασιολογικό Ιστό

Απαραίτητο μέσο για πλήθος εφαρμογών στον Σημασιολογικό ιστό, οι οντολογίες προορίζονται να προσφέρουν μία αναπαράσταση γνώσης και ένα λεξιλόγιο από κλάσεις και σχέσεις (κάτι αντίστοιχο με τους βιβλιοθηκονομικούς Θησαυρούς). Οι οντολογίες καταφέρνουν να συνενώσουν δύο ουσιώδη συστατικά, τα οποία συμβάλλουν στην ανάπτυξη του Παγκόσμιου Ιστού. Από τη μία ορίζουν την τυπική σημασιολογία της πληροφορίας διευκολύνοντας την επεξεργασία της από τον υπολογιστή, ενώ ταυτόχρονα από την άλλη ορίζουν σημασιολογία του πραγματικού κόσμου. Μ’ αυτό τον τρόπο επιτυγχάνεται η σύνδεση του περιεχομένου, το οποίο τυχαίνει μηχανικής επεξεργασίας, με τη σημασία που του δίνουν οι άνθρωποι βασιζόμενοι σε κοινά αποδεκτή ορολογία.

Οι οντολογίες μπορούν να διαδραματίσουν έναν πολύ σημαντικό ρόλο στις υπηρεσίες εξατομικευμένης ενημέρωσης, καθώς επιτρέπουν την περιγραφή, επεξεργασία και κατηγοριοποίηση του περιεχομένου τους, μ’ έναν τρόπο κοινά κατανοητό τόσο για τους χρήστες όσο και για τα συστήματα. Βασικό ενδιαφέρον των υπηρεσιών αυτών, άλλωστε, είναι η

όσο το δυνατόν καλύτερη εξυπηρέτηση των πελατών τους. Η ανάπτυξη, λοιπόν, ενός σημασιολογικά δομημένου συστήματος διευκολύνει την αναζήτηση και τελικά την πρόσβαση στις *αποθήκες* των ειδησεογραφικών τεκμηρίων, που αποτελούν μία πλούσια πηγή διάθεσης πληροφοριών.

Η παρούσα εργασία εστιάζει την μελέτη της κυρίως στην ανάπτυξη οντολογιών για το ειδησεογραφικό πεδίο (News Domain).

1.2 Στόχοι πτυχιακής εργασίας

Η παρούσα πτυχιακή εργασία πραγματοποιήθηκε με αφορμή την ανάγκη επέκτασης του Συστήματος Εξατομικευμένης Ενημέρωσης. Πρωταρχικός στόχος της πτυχιακής ήταν να βρεθούν ειδησεογραφικές πηγές, από τις οποίες να εξαγονται ειδήσεις για το Σύστημα Εξατομικευμένης Ενημέρωσης. Στη συνέχεια, στόχος ήταν η υλοποίηση των προγραμμάτων εξαγωγής πληροφορίας (wrappers), προκειμένου να εξάγονται τα νέα από της πηγές αυτές. Επίσης, στα πλαίσια της παρούσας εργασίας εντάσσεται και η μελέτη ανάπτυξης οντολογιών, ώστε στη συνέχεια να γίνει η κατηγοριοποίηση των ειδήσεων του συστήματος.

1.3 Διάρθρωση της πτυχιακής εργασίας

Στο **κεφάλαιο 2** επιχειρείται η παρουσίαση ορισμένων Συστημάτων εξατομικευμένης Ενημέρωσης, όπως για παράδειγμα είναι το *Mercurio.* Το **κεφάλαιο 3** περιλαμβάνει μία γενική περιγραφή του Συστήματος Εξατομικευμένης Ενημέρωσης πάνω στο οποίο διεξάχθηκε η παρούσα πτυχιακή εργασία. Το **κεφάλαιο 4** αποτελεί μία γενική μελέτη των οντολογιών, καθώς επίσης, παρουσιάζονται κάποια πρότυπα οντολογιών συγκεκριμένα για τον ειδησεογραφικό τομέα. Τέλος, στο **κεφάλαιο 5** παρουσιάζονται οι αλλαγές και οι προσθήκες, οι οποίες

πραγματοποιηθήκανε στο Σύστημα και συγκεκριμένα στη Βάση Δεδομένων Περιεχομένου (Database Context).

2. ΥΠΗΡΕΣΙΕΣ ΕΞΑΤΟΜΙΚΕΥΜΕΝΗΣ ΕΝΗΜΕΡΩΣΗΣ

Στην παρούσα ενότητα παρουσιάζονται υπηρεσίες εξατομικευμένης ενημέρωσης στον Ιστό. Αρχικά, αναφέρονται οι λόγοι για τους οποίους εφαρμόζονται τεχνικές εξατομίκευσης στην ειδησεογραφία και στη συνέχεια γίνεται μία αναφορά στις παραμέτρους που διαφοροποιούν τις υπηρεσίες εξατομικευμένης ενημέρωσης. Τέλος, παρουσιάζονται ορισμένες υπηρεσίες εξατομικευμένης ενημέρωσης.

2.1 Λόγοι ανάπτυξης υπηρεσιών εξατομικευμένης ενημέρωσης

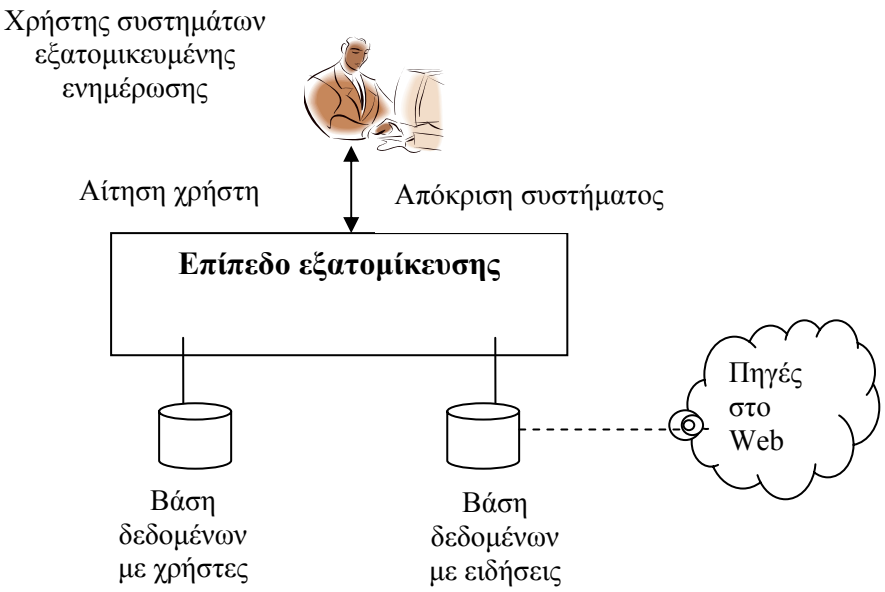
Ο Παγκόσμιος Ιστός ευνοεί ιδιαίτερα την ανάπτυξη επιχειρήσεων και εμπορικών συναλλαγών, καθώς αποτελεί ένα περιβάλλον εύχρηστο, γρήγορο και φθηνό. Γεγονός, το οποίο άλλωστε εύλογα υποδηλώνει η τεράστια δημοτικότητα του ηλεκτρονικού εμπορίου (*e-commerce*), των ηλεκτρονικών επιχειρήσεων (*e-business*), των ηλεκτρονικών τραπεζών (*e-banking*) και των ηλεκτρονικών εφημερίδων (*e-papers*). Μεγάλο ποσοστό αναγνωστών, πλέον, διαβάζει την εφημερίδα του ηλεκτρονικά. Διαρκώς, λοιπόν, αυξάνεται και ο αριθμός των εφημερίδων που δημοσιεύονται και σε ηλεκτρονική μορφή. Εκτός από τις εφημερίδες, ειδήσεις στον Παγκόσμιο Ιστό δημοσιεύονται από ειδησεογραφικά πρακτορεία, όπως το Reuters[1], από ενημερωτικά τηλεοπτικά μέσα, όπως το BBC[2], αλλά και από ειδησεογραφικά portals, όπως το in.gr[3]. Επίσης, ειδησεογραφικό υλικό προσφέρεται από υπηρεσίες εξατομικευμένης ενημέρωσης. Οι υπηρεσίες αυτές προσφέρονται από διάφορες εταιρείες ή ακόμα, από κάποιες ειδησεογραφικές πηγές.

[1] www.reuters.com
[2] www.bbc.co.uk
[3] www.in.gr/

Η ανάπτυξη μίας υπηρεσίας που προσφέρει εξατομικευμένη ενημέρωση έχει πολλά οφέλη, τόσο για τους χρήστες όσο και για την ίδια την εταιρεία. Όσον αφορά τους χρήστες, παρέχεται μεγάλη διευκόλυνση στην αναζήτηση πληροφοριών, καθώς στο Διαδίκτυο συνήθως οι τεράστιες ποσότητες δεδομένων δυσκολεύουν την αναζήτηση των ειδήσεων. Επιπλέον, οι εξατομικευμένες τεχνικές προσφέρουν στους χρήστες τη δυνατότητα αναζήτησης ανάλογα με τα ενδιαφέροντά τους. Ακόμα, οι υπηρεσίες αυτές προσθέτουν αξία στις υπηρεσίες τους με αποτέλεσμα να ικανοποιείται περισσότερο ο πελάτης, αυξάνοντας με τον τρόπο αυτό την επισκεψιμότητα των συγκεκριμένων υπηρεσιών, όπως και τη διάρκεια επίσκεψης.Όσο αυξάνεται η επισκεψιμότητα τόσο αυξάνονται και οι διαφημίσεις, άρα και τα έσοδα, καθώς οι διαφημίσεις αποτελούν το κυρίως έσοδο των υπηρεσιών αυτών.

Επίσης, η χρήση εξατομικευμένων υπηρεσιών στις διαφημίσεις αποτελεί ένα πάρα πολύ αποτελεσματικό τρόπο προώθησης προϊόντων. Οι υπηρεσίες ενημέρωσης άλλοτε προσφέρονται δωρεάν και άλλοτε απαιτείται κάποια συνδρομή.

Προτού προχωρήσουμε στην παρουσίαση των κατηγοριών εξατομίκευσης, γίνεται μία σύντομη αναφορά στην αρχιτεκτονική των συστημάτων εξατομικευμένης ενημέρωσης, η οποία φαίνεται σχηματικά και στην εικόνα 1. Ο χρήστης, λοιπόν, εισέρχεται στο σύστημα για να ενημερωθεί με τα σημερινά νέα. Η αίτησή του περνάει μέσα από το σύστημα εξατομικευμένης ενημέρωσης και ανάλογα με το προφίλ ανακτώνται και παρουσιάζονται οι ειδήσεις. Το προφίλ του χρήστη είναι αποθηκευμένο σε μία Βάση Δεδομένων με όλους τους εγγεγραμμένους χρήστες του συστήματος. Οι ειδήσεις είναι αποθηκευμένες στη Βάση Δεδομένων με τις ειδήσεις. Το σύστημα μπορεί να προσφέρει ειδήσεις από την ίδια την υπηρεσία ή από άλλες ειδησεογραφικές πηγές. Επίσης, το σύστημα μοντελοποιεί τους χρήστες και τις ειδήσεις ανάλογα με τις τεχνικές εξατομίκευσης και το επίπεδο λεπτομέρειας του συστήματος.



Εικόνα 1 : Αρχιτεκτονική σχεδίαση συστήματος εξατομικευμένης ενημέρωσης από το Web. [Αλεξόπουλος, 2003]

2.2 Κατηγοριοποίηση συστημάτων εξατομικευμένης ενημέρωσης

Στην ενότητα αυτή παρουσιάζονται οι συνηθισμένοι τρόποι εξατομίκευσης, σύμφωνα με τους οποίους στη συνέχεια θα γίνει η παρουσίαση και αξιολόγηση ορισμένων υπηρεσιών εξατομικευμένης ενημέρωσης.

Μία βασική διάκριση στα συστήματα εξατομικευμένης ενημέρωσης γίνεται με βάση τον τρόπο που συλλέγουν πληροφορίες για τον χρήστη. Το σύστημα μπορεί να δέχεται πληροφορίες για τα ενδιαφέροντα του χρήστη άμεσα, συνήθως μέσω κάποιας φόρμας που καλείται να συμπληρώσει ο χρήστης κατά την εγγραφή του στο σύστημα. Επίσης, η συλλογή των πληροφοριών αυτών μπορεί να πραγματοποιείται έμμεσα παρακολουθώντας τις επιλογές του χρήστη. Μία ακόμη πηγή πληροφοριών

είναι τα κληροδοτημένα δεδομένα (legacy data). Τέλος, πολύ συχνά τα συστήματα εφαρμόζουν συνδυασμό αυτών των μεθόδων.

Σύμφωνα με τα στοιχεία που συλλέγουν για το χρήστη, τα συστήματα μοντελοποιούν το χρήστη (user modeling), χτίζοντας το προφίλ του. Εκτός από τον χειρονακτικό τρόπο με τον οποίο κατασκευάζονται συνήθως τα μοντέλα των χρηστών, υπάρχουν, αυτόματοι τρόποι μοντελοποίησης που βασίζονται σε τεχνικές μηχανικής μάθησης.

Ακόμη, τα συστήματα εξατομικευμένης ενημέρωσης διακρίνονται σε προσαρμοστικά και μη. Μόλις το σύστημα συλλέξει τις απαραίτητες για την λειτουργία του πληροφορίες μπορεί είτε να προσαρμόζεται στις αλλαγές των προτιμήσεων του χρήστη είτε να παραμένει στατικό.

Τέλος, τα συστήματα κατηγοριοποιούνται σύμφωνα με τον τρόπο που φιλτράρουν και αναλύουν τα μεταδεδομένα του χρήστη ώστε να επιτευχθεί η εξατομίκευση. Διακρίνουμε, λοιπόν, το απλό φιλτράρισμα, το συνεργατικό φιλτράρισμα και το φιλτράρισμα βάση περιεχομένου. Ορισμένα σύστημα κάνουν συνδυασμό των παραπάνω.

2.3 Παρουσίαση εξατομικευμένων υπηρεσιών ενημέρωσης

Η παρούσα ενότητα περιέχει ορισμένες ειδησεογραφικές, και όχι μόνο, πηγές του Διαδικτύου, οι οποίες προσφέρουν υπηρεσίες εξατομικευμένης ενημέρωσης. Η αξιολόγησή τους γίνεται σύμφωνα με τις παραμέτρους που αναφέρονται στην προηγούμενη ενότητα.

Η *Washingtonpost*[4] στην ηλεκτρονική έκδοση προσφέρει υπηρεσίες εξατομίκευσης. Σ' ένα πρώτο επίπεδο η διεπαφή με τους χρήστες είναι κοινή, με έναν απλό προσωπικό χαιρετισμό. Οι υπηρεσίες εξατομίκευσης παρέχονται από την mywashingtonpost. Η συλλογή δεδομένων για το

[4] www.washingtonpost.com/

χρήστη είναι άμεση. Τα ενδιαφέροντα και οι προτιμήσεις του χρήστη δηλώνονται ρητά από τον χρήστη, επιλέγοντας τις κατηγορίες που τον ενδιαφέρουν περισσότερο. Εναλλακτικά, το σύστημα προτείνει υπερσυνδέσμους σχετικούς με τις μετοχές του χρηματιστηρίου που μπορεί να διατηρεί ο χρήστης (legacy data). Από κει και πέρα πρόκειται για ένα μη-προσαρμοστικό σύστημα και το φιλτράρισμα βασίζεται στο περιεχόμενο των άρθρων.

Εφημερίδες που χρησιμοποιούν παρόμοιο σύστημα είναι οι ηλεκτρονικές εκδόσεις της *Personal Wall Street Journal*[5] και της *San Francisco Chronicle*[6]*.* Η μόνη διαφορά είναι πως η *San Francisco Chronicle* επιτρέπει την παραμετροποίηση και εξατομίκευση και του τρόπου αναπαράστασης των ειδήσεων.

Το σύστημα *Mercurio* είναι ένα πρόγραμμα που εφαρμόζει η ισπανική εφημερίδα *ABC*[7] για την παροχή υπηρεσιών εξατομικευμένης ενημέρωσης. Σ᾽αυτό ο χρήστης πρέπει να δηλώσει ρητά τις προτιμήσεις του, οι οποίες χρησιμοποιούνται από το σύστημα για την κατασκευή του μοντέλου του χρήστη. Το σύστημα είναι προσαρμοστικό και λαμβάνει υπ’ όψη του τις κατηγορίες ειδήσεων που επιλέγει ο χρήστης και επίσης, ορισμένες λέξεις κλειδιά που υπάρχουν μέσα στο άρθρο της είδησης, τα οποία αποθηκεύονται στη Βάση για το προφίλ του χρήστη.

Μία άλλη υπηρεσία που προσφέρει υπηρεσίες εξατομικευμένης ενημέρωσης είναι το *yahoo*[8], μέσω της υπηρεσίας *my yahoo*[9]*.* Η συλλογή δεδομένων για το χρήστη πραγματοποιείται και εδώ σύμφωνα με τις επιλογές του χρήστη. Η διαφορά εδώ, όμως, είναι ότι οι ειδήσεις εξάγονται από ειδησεογραφικές πηγές που επιλέγει ο χρήστης. Η ανάκτηση πραγματοποιείται από πηγές οι οποίες παρέχουν RSS. Ο χρήστης, λοιπόν, καλείται να επιλέξει τα RSS feeds που τον ενδιαφέρουν και αυτά στη

[5] http://online.wsj.com/public/us
[6] www.sfgate.com/
[7] www.abc.es/
[8] www.yahoo.com
[9] http://my.yahoo.com/

συνέχεια παρουσιάζονται στην διεπαφή του *my yahoo.* Εκτός από τα RSS υποστηρίζεται εξατομίκευση και στις διαφημίσεις, στον καιρό, στις μετοχές, στα ζώδια, κόμικς και πολλά άλλα. Επίσης, παρέχεται στον χρήστη η δυνατότητα αλλαγής της εμφάνισης της διεπαφής. Το σύστημα προσαρμόζει το προφίλ του χρήστη ανάλογα με τα άρθρα που διαβάζει.

Η πλειοψηφία των συστημάτων εξατομικευμένης ενημέρωσης ταξινομούν τους υπερσυνδέσμους των κατηγοριών των άρθρων σύμφωνα με τις μέχρι τώρα προτιμήσεις των χρηστών. Αυτό συμβαίνει γιατί το σύστημα θεωρεί ότι ο χρήστης βρίσκει τον σύνδεσμο ενδιαφέρον (από την ανάγνωση του τίτλου του άρθρου), ακόμα και αν τελικά το ίδιο το άρθρο αποδειχτεί ότι δεν τον ενδιαφέρει. Όπως φάνηκε από την παραπάνω αξιολόγηση των συστημάτων, οι εφημερίδες έχουν περίπου έναν κοινό τρόπο εξατομίκευσης του περιεχομένου τους, ο οποίος δεν προωθεί τα προσαρμοστικά συστήματα.

3. ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΕΞΑΤΟΜΙΚΕΥΜΕΝΗΣ ΕΝΗΜΕΡΩΣΗΣ

Η παρούσα πτυχιακή εργασία διεξάγεται στα πλαίσια της ανάπτυξης μίας υπηρεσίας εξατομικευμένης ενημέρωσης, η οποία σ' ένα πρώτο στάδιο υλοποιήθηκε ως πτυχιακή εργασία του φοιτητή του Τμήματος Πληροφορικής και Τηλεπικοινωνιών Αλεξόπουλου Άγγελου και ως ανεξάρτητο έργο του Εργαστηρίου Τεχνολογίας Γνώσεων και Λογισμικού του Ε.Κ.Ε.Φ.Ε. Δημόκριτος. Η υπηρεσία εξατομικευμένης ενημέρωσης λειτουργεί αντλώντας αυτόματα πληροφορία από ειδησεογραφικές πηγές του Παγκόσμιου Ιστού. Αποτελείται από τον Content Server και από τον Personalization Server και προσφέρει τις παρακάτω λειτουργίες:

Συλλογή δεδομένων για τον χρήστη: Το σύστημα ανάλογα με τους υπερσυνδέσμους που επιλέγει ο χρήστης κατά την πλοήγηση του στην ιστοσελίδα κατασκευάζει το μοντέλο του χρήστη.
Μοντελοποίηση του χρήστη: Το σύστημα είναι προσαρμοστικό και ανάλογα με τις επιλογές του χρήστη ανανεώνει συνεχώς του μοντέλο του.
Φιλτράρισμα: Χρησιμοποιείται τόσο το φιλτράρισμα βάση περιεχομένου (λαμβάνοντας υπ' όψη τις κατηγορίες ειδήσεων και τις ειδησεογραφικές πηγές), όσο και το συνεργατικό φιλτράρισμα ομαδοποιώντας τους χρήστες σε ομάδες με κοινά χαρακτηριστικά. [Αλεξόπουλος, 2003]

3.1 Αρχιτεκτονική του συστήματος Content Server

Ο Content Server είναι υπεύθυνος για την πρόσβαση των χρηστών στις ειδήσεις που τους ενδιαφέρουν. Παίρνει είσοδο από τους χρήστες του συστήματος αλλά και από τις ειδησεογραφικές πηγές. Οι ειδησεογραφικές πηγές συμπεριλαμβάνουν υπηρεσίες του Παγκόσμιου Ιστού που δημοσιεύουν ειδήσεις (εφημερίδες, τηλεοπτικά μέσα, portals). Ο Content Server αφού επεξεργαστεί με τη Βοήθεια του Personalization Server τα στοιχεία που εισάγουν οι χρήστες στο σύστημα, εμφανίζει την προσωπική

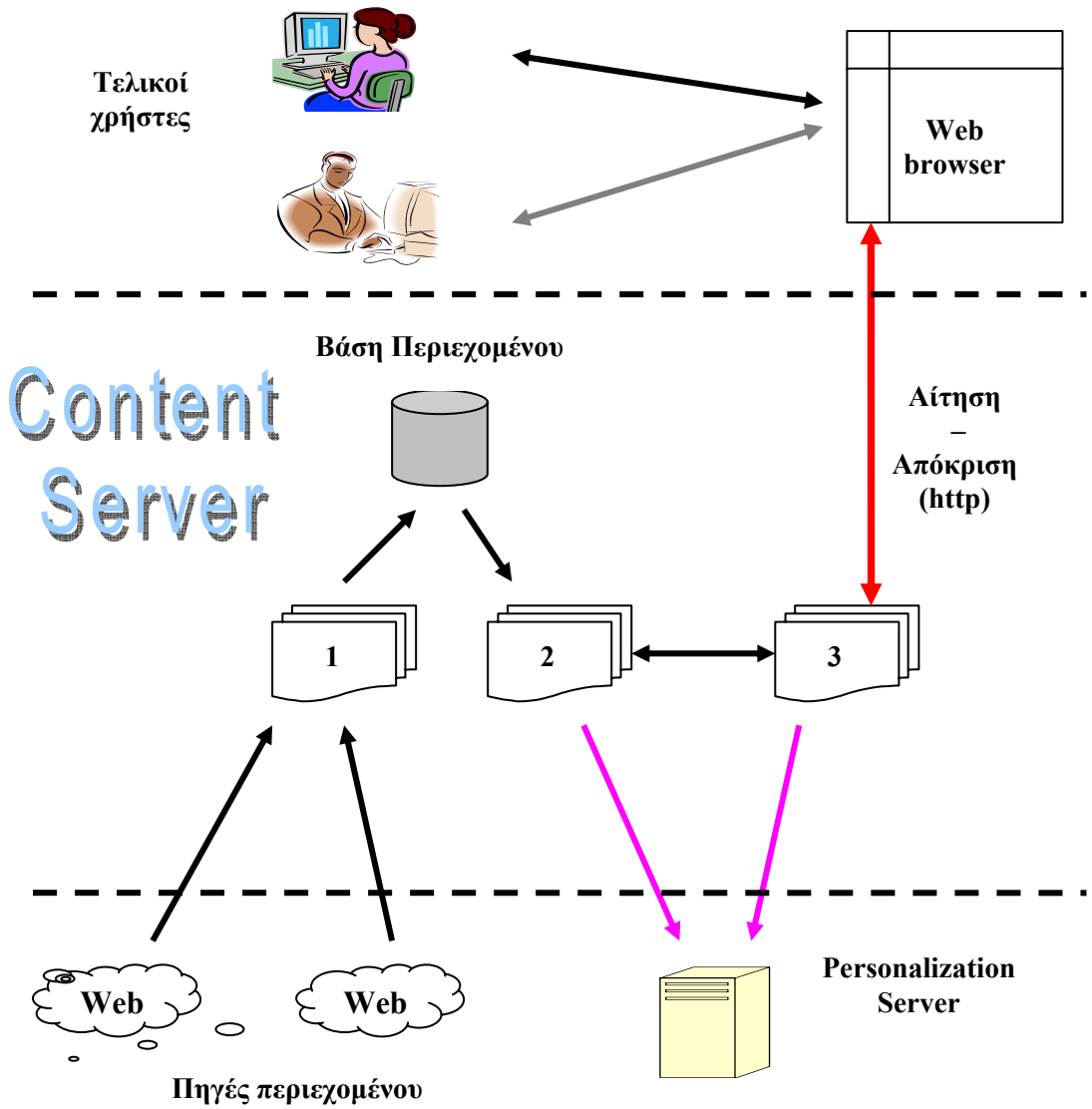
ηλεκτρονική εφημερίδα για τον συγκεκριμένο χρήστη με πρόσφατα άρθρα που ταιριάζουν στο προφίλ του. Στους νέους ή μη εγγεγραμμένους χρήστες το σύστημα εμφανίζει μία κοινή ηλεκτρονική εφημερίδα.

Κατά την εγγραφή τους στο σύστημα οι χρήστες καλούνται να συμπληρώσουν μία φόρμα με τα προσωπικά τους στοιχεία, τα οποία καθορίζουν ορισμένα χαρακτηριστικά τους (π.χ. φύλο, ηλικία), καθώς επίσης και να ορίσουν ένα όνομα χρήστη (username) και το συνθηματικό του (password). Σε συνδυασμό με τα στοιχειά που συλλέγει ο Content Server από τις επιλογές του χρήστη κατά την παραμονή του στην ιστοσελίδα, διαμορφώνεται και η προσωπική του ηλεκτρονική ιστοσελίδα.

Ο Content Server, όπως φαίνεται και στην **εικόνα 2** αποτελείται από:

1. Μονάδα Σάρωσης Περιεχομένου (Content Scanner Unit)
2. Μονάδα Επιλογής Περιεχομένου (Content Selector Unit)
3. Μονάδα Διεπαφής Περιεχομένου (Content Presenter Unit)

Σύμφωνα με την αρχιτεκτονική της **εικόνας 2** ο Content scanner παίρνει είσοδο από τους τελικούς χρήστες και από τις πηγές περιεχομένου (ειδησεογραφικές πηγές). Από τις πηγές περιεχομένου ανακτά ειδήσεις τις οποίες αποθηκεύει στη Βάση Περιεχομένου, όπου λόγω σεβασμού των πνευματικών δικαιωμάτων των πηγών αποθηκεύονται μόνο οι τίτλοι των ειδήσεων. Η Μονάδα Επιλογής αποφασίζει ποιές ειδήσεις θα παρουσιαστούν μέσω της Μονάδας Διεπαφής στο χρήστη. Επίσης, μέσω της Μονάδας Διεπαφής κατευθύνονται οι επιλογές των χρηστών στον Personalization Server. Οι πληροφορίες αυτές είναι απαραίτητες για την κατασκευή των προφίλ των χρηστών. Από τον Personalization Server λαμβάνει ο Επιλογέας περιεχομένου τα προφίλ των χρηστών προκειμένου να επιλέξει το κατάλληλο περιεχόμενο για την προσωπική εφημερίδα του χρήστη.

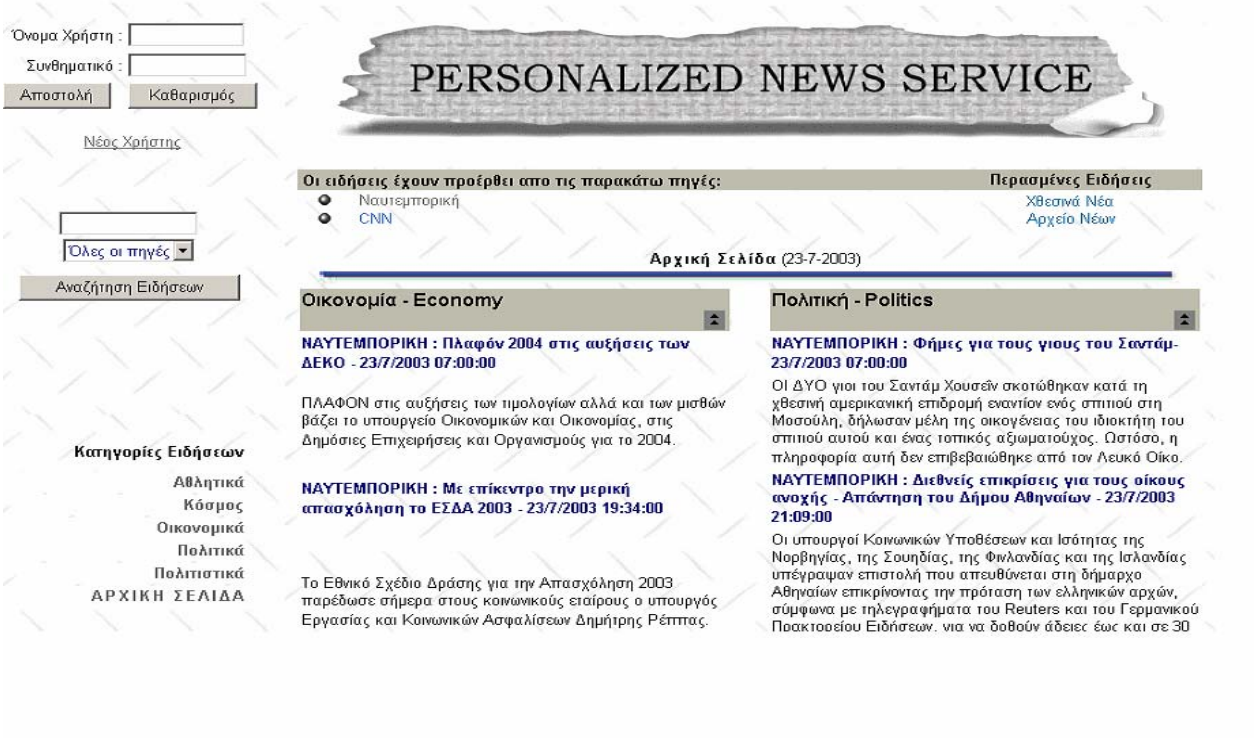


Εικόνα 1 : Η αρχιτεκτονική του Content Server
[Αλεξόπουλος, 2003]

3.1.1 Μονάδα Διεπαφής Περιεχομένου (Content Presenter)

Η Μονάδα Διεπαφής Περιεχομένου αναλαμβάνει την παρουσίαση των ειδήσεων που έχουν επιλεχθεί από τον Content Selector κατά τη διάρκεια της εξατομίκευσης. Όλες οι λειτουργίες της υπηρεσίας προσφέρονται στους χρήστες μέσω του Content Presenter. Οι λειτουργίες που προσφέρονται είναι οι εξής:

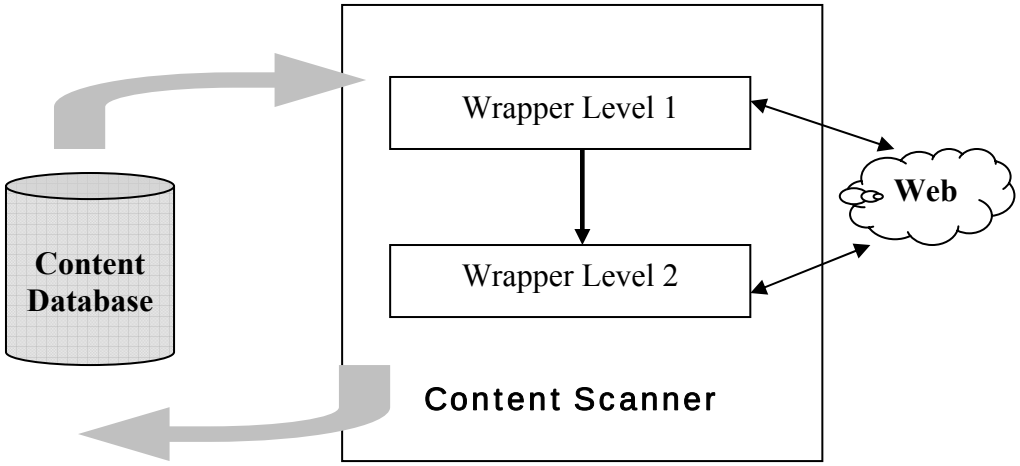
- Παρουσίαση των πιο πρόσφατων ειδήσεων που υπάρχουν στην Βάση Περιεχομένου του Content Server δίχως εξατομίκευση.
- Εγγραφή νέων χρηστών και πιστοποίηση της ταυτότητας των εγγεγραμμένων χρηστών.
- Παρουσίαση των πιο πρόσφατων ειδήσεων που υπάρχουν στην Βάση Περιεχομένου της εφαρμογής με βάση τις προτιμήσεις άλλων χρηστών (προσωπική ηλεκτρονική εφημερίδα).
- Παρουσίαση των πιο πρόσφατων ειδήσεων που υπάρχουν στην Βάση Περιεχομένου της εφαρμογής με βάση τις προτιμήσεις άλλων χρηστών με παρόμοια χαρακτηριστικά.
- Αναζήτηση και παρουσίαση ειδήσεων που περιέχουν κάποια λέξη – κλειδί στον τίτλο ή στην πρώτη φράση του άρθρου.
- Παρουσίαση των ειδήσεων της προηγούμενης ημέρας με εξατομικευμένο τρόπο.
- Αναζήτηση ειδήσεων με βάση την ημερομηνία, την κατηγορία και την πηγή από το "Αρχείο Ειδήσεων", όπου φυλάσσονται όλες οι ειδήσεις που έχει εξάγει το σύστημα από τον Παγκόσμιο Ιστό.



Εικόνα 3: Γενική μορφή της Μονάδας Διεπαφής Περιεχομένου (χωρίς εξατομικευμένες υπηρεσίες). Η ιστοσελίδα χωρίζεται σε 2 άνισα πλαίσια. [Άλεξόπουλος, 2003]

3.1.2 Σαρωτής Περιεχομένου (Content Scanner)

Ο Content Scanner είναι υπεύθυνος για την εξαγωγή των ειδήσεων από τις ειδησεογραφικές πηγές του Ιστού. Για την διαδικασία εξαγωγής ειδήσεων αναπτύχθηκαν προγράμματα εξαγωγής πληροφορίας (wrappers). Οι wrappers ανά τακτά χρονικά διαστήματα ανιχνεύουν τις πηγές προκειμένου να αντλήσουν καινούριες ειδήσεις. Μόλις ανανεωθεί το περιεχόμενο των ειδήσεων ο Scanner στέλνει τις νέες ειδήσεις στη Βάση Περιεχομένου (Content Database), με σκοπό να χρησιμοποιηθούν στη διαδικασία εξατομίκευσης. Η επιλογή βασίζεται στο προφίλ του χρήστη όπως αυτό διαμορφώνεται κατά την αλληλεπίδρασή του με το σύστημα. Σχηματικά η αρχιτεκτονική δομή του Content Scanner παρουσιάζεται στην **εικόνα 4**.



Εικόνα 4: Η δομή του Content Scanner
[Αλεξόπουλος, 2003]

Ως είσοδο ο Content Scanner παίρνει ορισμένες παραμέτρους που έχουν τεθεί στην Βάση Δεδομένων Περιεχομένου, όπως ποιές πηγές θα χρησιμοποιήσει ή πόσες ειδήσεις θα αντλήσει και ως έξοδο επιστρέφει τις ειδήσεις στη Βάση Δεδομένων Περιεχομένου. Οι wrappers λειτουργούν χρησιμοποιώντας κανονικές εκφράσεις (regular expressions), οι οποίες είναι αποθηκευμένες στη Βάση Περιεχομένου και μπορούν εύκολα να αλλάξουν από τους χρήστες του συστήματος, γεγονός το οποίο ενδείκνυται

όταν αλλάζει η δομή παρουσίασης των ειδήσεων από κάποια ειδησεογραφική πηγή. Η διαδικασία εξαγωγής των ειδήσεων παρουσιάζεται πιο αναλυτικά στο κεφάλαιο 5.

3.2 Εξυπηρετητής Εξατομίκευσης (Personalization Server)

Εκτός από τον Content Server, η υπηρεσία εξατομικευμένης ενημέρωσης βασίζεται επίσης στον Personalization Server, ο οποίος αναπτύχθηκε σε ανεξάρτητο έργο του εργαστηρίου Τεχνολογίας Γνώσεων και Λογισμικού του Ε.Κ.Ε.Φ.Ε. Δημόκριτος. Ο Personalization Server κατασκευάζει και διατηρεί τα προφίλ των χρηστών, τα οποία ο Content Selector χρησιμοποιεί για την εξατομίκευση του περιεχομένου που παρουσιάζει στον τελικό χρήστη. Ως είσοδο παίρνει ορισμένες πληροφορίες (άλλες υποχρεωτικές, άλλες προαιρετικές) που μπορεί να δώσει ο χρήστης κατά την εγγραφή του στο σύστημα, όπως προσωπικά στοιχεία, κάποιες προτιμήσεις, κτλ. Επίσης, αποθηκεύεται και η συμπεριφορά του χρήστη κατά την πλοήγηση, όπως ποιές κατηγορίες ειδήσεων επιλέγει να διαβάσει. Η έξοδος του Personalization Server είναι το προφίλ του χρήστη.

4. ΟΝΤΟΛΟΓΙΕΣ

Στις μέρες μας παρατηρείται αυξημένη η χρήση των υπηρεσιών εξατομικευμένης ενημέρωσης από τον Ιστό, οι οποίες προσπαθούν να ικανοποιήσουν τις ολοένα και αυξανόμενες απαιτήσεις των χρηστών. Παρόμοιες προκλήσεις, βέβαια, έχουν να αντιμετωπίσουν γενικότερα οι ηλεκτρονικές υπηρεσίες που προσφέρουν ειδησεογραφική, και όχι μόνο, ενημέρωση. Προκειμένου, λοιπόν, να προσφέρεται καλύτερη εξυπηρέτηση στους πελάτες τους οι εταιρείες κατευθύνονται προς την ανάπτυξη ενός σημασιολογικά δομημένου συστήματος, ώστε να επιτύχουν αποδοτική αναζήτηση και επεξεργασία δεδομένων. Προς αυτή την κατεύθυνση κινούνται και οι οντολογίες.

4.1 Βασική παρουσίαση των οντολογιών

Ο όρος «οντολογία» έχει μακρά ιστορία που προέρχεται από τη φιλοσοφία και αναφέρεται στον κλάδο αυτό της μεταφυσικής που ασχολείται με τη φύση και την οργάνωση των όντων. Η επιστήμη και η Τεχνητή Νοημοσύνη υιοθέτησαν τον όρο αυτό για να δηλώσουν «μία διαμοιρασμένη και κοινή κατανόηση κάποιου τομέα, η οποία μπορεί να ανταλλαγεί μεταξύ ανθρώπων και συστημάτων εφαρμογών» [Gruber].

Για την επιστήμη οι οντολογίες είναι εργαλεία αναπαράστασης (representation) και συλλογιστικής (reasoning). Εκτενέστερα αλλά και πιο συγκεκριμένα, η οντολογία ορίζεται ως μία τυπική (formal), κατηγορηματική (explicit) προδιαγραφή μίας διαμοιρασμένης (shared) εννοιολογικής αναπαράστασης (conceptualization). Ο όρος *εννοιολογική αναπαράσταση* αναφέρεται σε ένα αφηρημένο μοντέλο φαινομένων του κόσμου στον οποίο έχουν προσδιοριστεί οι έννοιες που σχετίζονται με τα φαινόμενα αυτά. Ο όρος *κατηγορηματική* σημαίνει ότι το είδος των εννοιών που χρησιμοποιούνται, και οι περιορισμοί που αφορούν τη χρήση των εννοιών αυτών είναι προσδιορισμένα με σαφήνεια. Ο όρος *τυπική* αναφέρεται στο ότι η οντολογία πρέπει να είναι μηχανικά αναγνώσιμη. Ο

όρος *διαμοιρασμένη* αναφέρεται στην γνώση κοινής αποδοχής που πρέπει να αποτυπώνει η οντολογία στα πλαίσια μίας κοινότητας. [Gruber – από τις σημειώσεις του μαθήματος *Διαχείριση Γνώσης* των καθηγητών Μ. Γεργατσούλη και Χ. Παπαθεοδώρου του Τμήματος Αρχειονομίας - Βιβλιοθηκονομίας].

Με την ανάπτυξη των οντολογιών γίνεται μία προσπάθεια να περιγράφει ένα πεδίο γνώσης (Knowledge Domain), ώστε να υπάρχει ένας κοινός τρόπος επικοινωνίας μεταξύ εκείνων που δραστηριοποιούνται στο συγκεκριμένο πεδίο. Επιπλέον, χρειάζεται να είναι αυστηρά θεμελιωμένη (formal model), ώστε να είναι μηχανικά κατανοητή και επεξεργάσιμη (machine readable).

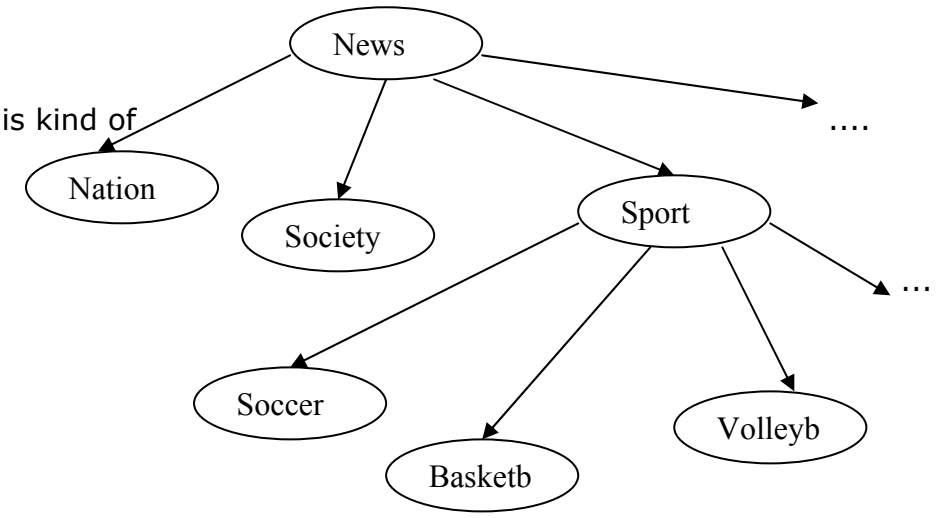
Οι οντολογίες αποτελούν βασική πηγή για τις εφαρμογές στο Σημασιολογικό Ιστό, και όχι μόνο, καθώς παρέχουν μία γλώσσα αναπαράστασης γνώσης και ένα λεξικό από *κλάσεις (classes)* και *σχέσεις (relations)*, τις οποίες οι διάφορες υπηρεσίες μπορούν να χρησιμοποιήσουν για να περιγράψουν το περιεχόμενό τους και να το επεξεργαστούν κατάλληλα.

Τί όμως τελικά είναι οι οντολογίες και ποίος είναι ο ρόλος των *κλάσεων* και των σχέσεων; Μία οντολογία αποτελείται από την ταξινομία (taxonomy), η οποία και την χαρακτηρίζει, και από ένα σύνολο συμπερασματικών κανόνων. Η ταξινομία καθορίζει τις κλάσεις των αντικειμένων και τις μεταξύ τους σχέσεις, ενώ οι συμπερασματικοί κανόνες αναπαριστούν τη λογική. Για παράδειγμα, η πρόταση «η *παθολογία* αποτελεί κλάδο της *Ιατρικής*» αποτελεί έναν συμπερασματικό κανόνα, ο οποίος απορρέει ως συμπέρασμα μίας ταξινομίας, στην οποία η παθολογία ορίζεται ως υποκατηγορία της Ιατρικής. Επιπλέον, οι όροι "παθολογία" και "Ιατρική" θεωρούνται κλάσεις του πεδίου «Ιατρική». Οι κλάσεις μαζί με τις παρακάτω κατηγορίες αποτελούν τα βασικά συστατικά μίας οντολογίας.

Μία οντολογία περιέχει πέντε βασικές κατηγορίες συστατικών:

- Κλάσεις (Classes): Οι κλάσεις, όπως είδαμε προηγουμένως είναι έννοιες που σχετίζονται με ένα πεδίο. Για παράδειγμα, σε μία οντολογία που αφορά το πεδίο των ειδήσεων η «πολιτική» ή τα «αθλητικά» αποτελούν κλάσεις.
- Σχέσεις (relations): Ένας τύπος αλληλεπίδρασης μεταξύ των εννοιών ενός πεδίου. Όπως για παράδειγμα: Subclass–of, is – a και άλλα.
- Συναρτήσεις (functions): Μία περίπτωση σχέσης κατά την οποία το v-οστό στοιχείο της σχέσης προσδιορίζεται μοναδικά από τα v-1 προηγούμενα στοιχεία.
- Αξιώματα (axioms): αναπαριστούν προτάσεις, οι οποίες είναι αληθείς. Για παράδειγμα, αν ο φοιτητής Β. παρακολουθεί το $2^{ο}$ έτος της σχολής τότε μπορεί να εγγραφεί στο μάθημα Ψ.
- Στιγμιότυπα (instances): Αναπαριστούν συγκεκριμένα στοιχεία. Για παράδειγμα, ο φοιτητής με το όνομα Νίκος είναι ένα στιγμιότυπο της κλάσης «φοιτητής».

Η παρακάτω εικόνα (**εικόνα 5**) αποτελεί έναν απλοποιημένο σχεδιασμό μίας οντολογίας που περιέχει κλάσεις και σχέσεις.



Εικόνα 5: Σχεδιάγραμμα διακλάδωσης μίας οντολογίας ειδησεογραγικού περιεχομένου [Jianqing L., Liu Shengping, Lin ZuoQuan, Wu Cen]

Οι οντολογίες ανάλογα με το είδος της πληροφορίας που καλούνται κάθε φορά να αναπαραστήσουν χωρίζονται σε ορισμένες κατηγορίες. Ορισμένες από τις πιο βασικές περιγράφονται παρακάτω[Γεργατσούλης Μ., Χ. Παπαθεοδώρου]:

- *Γενικές ή κοινές οντολογίες* (generic ή common sense ontologies ή Top-level ontologies): Περιγράφουν πολύ γενικές έννοιες (όπως ο χρόνος, ο χώρος, το γεγονός), οι οποίες είναι ανεξάρτητες από ένα συγκεκριμένο πρόβλημα ή περιοχή γνώσης.
- *Οντολογίες πεδίου ορισμού* (Domain Ontologies): Είναι οι οντολογίες εκείνες που σκοπό έχουν να αναπαραστήσουν γνώση γύρω από ένα συγκεκριμένο πεδίο, όπως η ιατρική, η ειδησεογραφία κλπ.
- *Οντολογίες μεθοδολογίας ή εργασιών* (method or task ontologies): περιγράφουν λεξιλόγιο για μία γενική δραστηριότητα ή εργασία, π.χ. διάγνωση.
- *Οντολογίες μεταδεδομένων* (metadata ontologies): παρέχουν ορολογία για την περιγραφή πληροφορίας, η οποία διατίθεται ηλεκτρονικά. Η ορολογία μπορεί να περιέχει μεταδεδομένα για συγκεκριμένο πεδίο γνώσης.
- *Οντολογίες αναπαράστασης* (representational ontologies): παρέχουν οντότητες αναπαράστασης χωρίς να προσδιορίζουν τι συγκεκριμένο αναπαριστούν. Για παράδειγμα η *Frame ontology* ορίζει έννοιες όπως frames, slots κλπ.
- *Οντολογίες εφαρμογής* (application ontologies): Είναι οι πιο συγκεκριμένες οντολογίες, που σκοπό έχουν να καλύψουν τις ανάγκες μίας συγκεκριμένης εφαρμογής.

Οι οντολογίες προσφέρουν ένα εννοιολογικό μοντέλο ενός πεδίου ενδιαφέροντος καθιστώντας τη γνώση επαναχρησιμοποιήσιμη και διαμοιραζόμενη. Λόγω, λοιπόν, της πρακτικότητας και της δυναμικότητάς τους, οι οντολογίες έχουν καταφέρει να γίνουν δημοφιλείς σε πολλές ερευνητικές ομάδες και επιχειρησιακές κοινότητες και να αποτελέσουν ακρογωνιαίο λίθο σε πολλές εφαρμογές. Ενδεικτικά αναφέρονται τα παρακάτω:

- Συστήματα διαχείρισης γνώσης (Knowledge management), όπου χρησιμοποιούνται οντολογίες με σκοπό τη μοντελοποίηση της γνώσης και τη «μετάφραση της πληροφορίας».
- Ηλεκτρονικό εμπόριο (e-commerce), για την ανάκτηση γνώσης και πληροφορίας. Για τη διαδικασία αυτή οι οντολογίες χρησιμοποιούνται από πράκτορες λογισμικού (Software agents). Οι πράκτορες λογισμικού είναι προγράμματα που εκτελούν κάποια λειτουργία και με το πέρας της εκτέλεσης αυτής παράγουν αποτελέσματα. Συνήθως, επισκέπτονται ιστοσελίδες και επεξεργάζονται τις πληροφορίες που βρίσκουν σε αυτές που επισκέπτονται. Οι λειτουργίες για τις οποίες χρησιμοποιούνται είναι διάφορες, όπως η εύρεση, ταξινόμηση και επιλογή δεδομένων. Στο ηλεκτρονικό εμπόριο εκτελούν λειτουργίες όπως η σύγκριση τιμών του ίδιου προϊόντος σε διάφορα ηλεκτρονικά καταστήματα.
- Διαλειτουργικότητα (interoperability) μεταξύ συστημάτων: Οι οντολογίες προσφέρουν μία κοινή γλώσσα σε διάφορους χρήστες για εποικοδομητική ανταλλαγή δεδομένων ή χρησιμοποιούνται ως μεταφραστές διαφορετικών πακέτων λογισμικού. Επιπλέον, οι οντολογίες χρησιμοποιούνται για την υποστήριξη μετάφρασης μεταξύ διαφορετικών γλωσσών και αναπαραστάσεων.
- Ψηφιακές Βιβλιοθήκες (Digital libraries), για καλύτερη ταξινόμηση του υλικού τους.
- Στα πολυμέσα (Multimedia) και στις Τηλεπικοινωνίες (Telecommunications).

4.2 Λόγοι ανάπτυξης οντολογιών

Τα τελευταία χρόνια η ανάπτυξη των οντολογιών έχει αποτελέσει αντικείμενο ενδιαφέροντος σ’ ένα ευρύ πεδίο ερευνητών. Στο Διαδίκτυο μπορεί να βρεί κανείς διάφορες οντολογίες, από μεγάλες ταξινομίες κατηγοριοποίησης Ιστότοπων, όπως το yahoo, μέχρι κατηγοριοποιήσεις προϊόντων προς πώληση (όπως στο Amazon). Γιατί κρίνεται όμως αναγκαία η ανάπτυξη και τελικά η εφαρμογή οντολογιών; Παρακάτω γίνεται μία

προσπάθεια να μελετηθούν οι λόγοι που υποδεικνύουν την σκοπιμότητα των οντολογιών.

Τον πιο διαδεδομένο λόγο δημιουργίας οντολογιών αποτελεί η *διαμοιρασμένη και κοινή κατανόηση* της δόμησης της πληροφορίας τόσο ανάμεσα στους ανθρώπους όσο και ανάμεσα στους πράκτορες λογισμικού (Software agents) [Musen 1992; Gruber 1993]. Για παράδειγμα, διαφορετικές ιστοσελίδες που παρέχουν ιατρικές πληροφορίες ή προωθούν ιατρικές υπηρεσίες ηλεκτρονικού εμπορίου (e-commerce) μπορεί να μοιράζονται και να εκδίδουν κοινή οντολογία με όλους τους όρους που χρησιμοποιούν. Μ' αυτό τον τρόπο, οι πράκτορες λογισμικού μπορούν να εξάγουν και να συλλέξουν πληροφορίες από όλες αυτές τις διαφορετικές ιστοσελίδες.

Η δυνατότητα επαναχρησιμοποίησης της γνώσης ενός συγκεκριμένου πεδίου (domain knowledge) συντέλεσε σε μεγάλο βαθμό στην ανάπτυξη των οντολογιών. Για παράδειγμα, μοντέλα για πολλά διαφορετικά πεδία γνώσης χρησιμεύουν για την αναπαράσταση του *χρόνου,* π.χ. ώρες αφίξεων, σχετικές μετρήσεις του χρόνου κλπ. Αν, λοιπόν, αναπτυχθεί από μία ομάδα ερευνητών μία σχετική λεπτομερής οντολογία, μπορεί να επαναχρησιμοποιηθεί από κάποιους άλλους, οι οποίοι μπορεί να την χρειάζονται για διαφορετικού είδους υπηρεσίες. Επιπλέον, για την δημιουργία μία μεγάλης οντολογίας, μπορούν να χρησιμοποιηθούν ήδη υπάρχουσες οντολογίες που περιγράφουν μέρη ενός πεδίου και να ενοποιηθούν. Επίσης, γενικές οντολογίες μπορούν να επεκταθούν, προκειμένου να περιγράψει κανείς συγκεκριμένα πεδία.

Ακόμα, μπορεί κάποιος να αναπτύξει μία οντολογία προκειμένου να ορίσει ρητά (explicit) τις υποθέσεις που ισχύουν σε ένα επιστημονικό πεδίο. Με τον τρόπο αυτό, σε περίπτωση που αλλάξει η γνώση μας για ένα πεδίο μπορούν εύκολα να πραγματοποιηθούν οι απαραίτητες αλλαγές.

Μία ακόμα ευρεία χρήση των οντολογιών συνιστά ο διαχωρισμός του πεδίου γνώσης (domain knowledge) από τη λειτουργική γνώση

(operational knowledge). Μ'αυτό τον τρόπο, μπορούμε να περιγράψουμε τη διαδικασία διαμόρφωσης (configuring) ενός προϊόντος από τα συστατικά του σύμφωνα με την απαιτούμενη τεχνική περιγραφή.

Επίσης, η ανάλυση γνώσης συγκεκριμένου πεδίου (Domain Knowledge) γίνεται δυνατή, καθώς διατίθεται μία προτυποποιημένη περιγραφή των όρων. Η τυπική ανάλυση των όρων είναι ιδιαίτερα χρήσιμη σε περίπτωση είτε επαναχρησιμοποίησης ήδη υπαρχόντων οντολογιών είτε επέκτασης κάποιων άλλων γενικών οντολογιών.

4.3 Γλώσσες αναπαράστασης οντολογιών

Οι οντολογίες εκφράζονται συνήθως σε μία γλώσσα βασισμένη στη λογική, έτσι ώστε να μπορούν να γίνουν λεπτομερείς, ακριβείς, ορθές και εκφραστικές διακρίσεις μεταξύ των τάξεων, των ιδιοτήτων και των σχέσεων της οντολογίας. Κατά καιρούς έχουν αναπτυχθεί διάφορες γλώσσες για την αναπαράσταση οντολογιών, από τις οποίες άλλες χαρακτηρίζονται ως παραδοσιακές γλώσσες (όπως η ontolingua[10]) και άλλες ως web-based γλώσσες και υπάρχουν επίσης γλώσσες που αναπτύχθηκαν για να αναπαραστήσουν συγκεκριμένες οντολογίες και να χρησιμοποιηθούν σε συγκεκριμένες εφαρμογές (CycL, GRAIL, NKRL). Η διαφορά ανάμεσα στις παραδοσιακές και στις web-based γλώσσες είναι πως οι τελευταίες διαθέτουν καλά ορισμένη σύνταξη και σημασιολογία και ικανοποιητική συλλογιστική (reasoning) υποστήριξη. Επίσης, παρέχουν δύναμη και ευελιξία στην εκφραστικότητα και το συντακτικό τους είναι συμβατό με ήδη υπάρχοντα πρότυπα του web (XML, RDF, RDFS).

Συλλογιστική (reasoning) ονομάζεται η διαδικασία επαγωγής συμπερασμάτων με τη χρήση της λογικής και εξακρίβωσης της εγκυρότητας των συμπερασμάτων αυτών. Η συλλογιστική υποστήριξη (υπηρεσίες συμπερασμού) εξασφαλίζει ποιότητα στην οντολογία. Ποιότητα

[10] www.ksl.stanford.edu/software/ontolingua/

στην οντολογία επιτυγχάνεται, επίσης, και με την ανάπτυξη μίας γλώσσας, η οποία διαθέτει πλούσια εκφραστικότητα. Ο όρος *εκφραστικότητα* αναφέρεται στον αριθμό και στον τύπο των διαθέσιμων σχέσεων που ορίζουν τις κλάσεις.

Οι οντολογικές γλώσσες κατά κόρον βασίζονται στην XML (eXtensible Markup Language) και στην RDF (Resource Description Framework), γλώσσες καθιερωμένες στον Παγκόσμιο Ιστό. Η XML από μόνη της δεν μπορεί να χρησιμοποιηθεί για δημιουργία οντολογιών εξειδικευμένου πεδίου ή οντολογικών λεξιλογίων και δεν μπορεί να χρησιμοποιήσει βασικές οντολογικές αρχές μοντελοποίησης, καθώς το μοντέλο δεδομένων της είναι χαμηλού επιπέδου και δεν διαθέτει μηχανή συμπερασματολογίας. Η RDF συστάθηκε από τη W3C (WWW Consortium[11]) και προορίζονταν αρχικά για την αναπαράσταση μεταδεδομένων. Διαθέτει ένα τυποποιημένο μοντέλο δεδομένων και ένα τυποποιημένο XMLs συντακτικό. Σε αντίθεση με την XML επιτρέπει την αναπαράσταση *ορισμένης* οντολογικής γνώσης. Από την άλλη, όμως, δεν ανταποκρίνεται στην ανάπτυξη μίας υψηλής σημασιολογικής αναπαράστασης και αδυνατεί να περιγράψει εις βάθος το νόημα της πληροφορίας.

Επέκταση της RDF, το RDF Schema είναι μία γλώσσα με την οποία το μοντέλο δεδομένων της RDF εμπλουτίζεται με χαρακτηριστικά αντικειμενοστραφούς αναπαράστασης, όπου ο πόρος αντιστοιχεί σε αντικείμενο. Συγκεκριμένα, το RDF Schema ορίζει ένα λεξικό για να εκφράζονται οι κλάσεις των πόρων, οι πόροι, οι ιδιότητες του και οι μεταξύ τους σχέσεις. Πόρος είναι οτιδήποτε θέλουμε να περιγράψουμε, για παράδειγμα μία ιστοσελίδα, ένας δικτυακός τόπος, μία έννοια κτλ.

Μία γλώσσα ακόμα που στήριξε το συντακτικό της στην RDF Schema, η DAML (DAPRA Agent Markup Language) +OIL (Ontology Inference Layer) αναπτύχθηκε από την US Defense Advanced Research Project Agency (DAPRA[12]) σε συνεργασία με την EU committee on agent markup

[11] http://www.w3.org/
[12] www.dapra.com

languages. Όπως κάθε οντολογική γλώσσα, η DAML+OIL σχεδιάστηκε για να περιγράψει τη δομή (structure) ενός πεδίου γνώσης. Διατηρώντας αντικειμενική προσέγγιση η δομή περιγράφεται με τάξεις και ιδιότητες. Μία DAML +OIL οντολογία περιέχει ένα σύνολο από *αξιώματα,* τα οποία υποστηρίζουν χαρακτηριστικά των τάξεων και των ιδιοτητών αυτών.

Η OWL (Web Ontology Language: www.w3.org/TR/owl-features/) σχεδιάστηκε από την W3C με σκοπό να αποτελέσει πιο πλούσια - από άποψη εκφραστικότητας - προέκταση της RDF. Όσο μεγαλύτερη, όμως, εκφραστικότητα χαρακτηρίζει μία γλώσσα, τόσο πιο αναποτελεσματική αποδεικνύεται η συλλογιστική της υποστήριξη. Ως λύση στο πρόβλημα που ανακύπτει η ομάδα εργασίας του W3C αποφάσισε να προχωρήσει στην ανάπτυξη τριών διαφορετικών «*υπο-γλωσσών*» (sublanguages), οι οποίες θα συγκροτούν την OWL: η OWL Lite, η OWL DL (Description Logic) και η OWL Full. Η καθεμία από αυτές προσαρμόζεται έτσι ώστε να καλύπτονται διαφορετικές ανάγκες.

Η OWL Full αποτελεί ολόκληρη την OWL περιλαμβάνοντας επίσης, την OWL Lite και την OWL DL. Βρίσκεται σε πλήρη συμβατότητα με την RDF, η συντακτική ελευθερία της οποίας, προσφέρει στην γλώσσα πολύ υψηλή εκφραστικότητα. Μ’ αυτό τον τρόπο, όμως, η γλώσσα επεκτείνεται χωρίς περιορισμούς δυσκολεύοντας την όποια ολοκληρωμένη συλλογιστική υποστήριξη. Η OWL DL από την άλλη παρέχει μία σαφώς ορισμένη σημασιολογία χωρίς να στερείται εκφραστικότητας, αλλά και με ικανοποιητική δυνατότητα εξαγωγής συμπερασμάτων. Τέλος, η OWL Lite διαθέτει μία ιεραρχία κλάσεων με απλούς περιορισμούς και για το λόγο αυτό γίνεται εύκολα κατανοητή από τους χρήστες και εύκολα επεξεργάσιμη από τους υπολογιστές. Από την άλλη, όμως, έχει το μειονέκτημα περιορισμένης εκφραστικότητας.

Πέρα από την OWL και την DAML +OIL υπάρχουν πολλές άλλες γλώσσες για την αναπαράσταση οντολογιών, οι οποίες ως επί των πλείστων βασίζονται στην XML, RDF και RDF Schema. Αναφέρουμε επιγραμματικά

μερικές, όπως η Simple HTML Ontology Extensions (SHOE[13]), η Ontology Exchange Language (XOL[14]), η Ontology Markup language [15](OML και KML), η Riboweb[16] και πολλές ακόμα.

4.4 Οντολογίες και πρότυπα για τον ειδησεογραφικό τομέα

Ένας χώρος, ο οποίος φαίνεται να έλκεται ιδιαίτερα προσκομίζοντας οφέλη από την ανάπτυξη των οντολογιών, είναι αυτός της ειδησεογραφίας. Στις σύγχρονες κοινωνίες, όπου οι άνθρωποι είτε για προσωπικούς είτε για επαγγελματικούς λόγους αναζητούν διαρκώς τη γνώση και η πληροφόρηση αποτελεί καθημερινή επιδίωξη, οι εφημερίδες και γενικότερα τα μέσα μαζικής ενημέρωσης χαίρουν εκτίμησης. Αναγκασμένες, λοιπόν, οι ειδησεογραφικές υπηρεσίες από τις απαιτήσεις των καιρών στρέφονται προς τις νέες τεχνολογίες που υποδεικνύουν την ηλεκτρονική δημοσίευση ως την πλέον κατάλληλη προκειμένου να ικανοποιηθούν οι ολοένα και αυξανόμενες ανάγκες των πελατών τους.

Μέσα σ' όλη αυτή την κινητικότητα για την ανάπτυξη εργαλείων, που θα εφαρμόζουν τις νέες τεχνολογίες του Σημασιολογικού Ιστού στον κόσμο της ενημέρωσης, πραγματοποιήθηκαν αρκετές προσπάθειες ανάπτυξης οντολογιών για το συγκεκριμένο πεδίο. Καθώς τα πεδία εφαρμογής των οντολογιών διευρύνονται, καθίσταται φανερό ότι μία οντολογία γενικού περιεχομένου (general ontology) δεν θα μπορούσε να αντεπεξέλθει στις ανάγκες διαφορετικών περιεχομένου περιοχών.

[13] www.cs.umd.edu/projects/plus/SHOE/
[14] www.ai.sri.com/pkarp/xol/
[15] http://xml.coverpages.org/oml9808.html
[16] http://www.cs.man.ac.uk/~stevensr/onto/node6.html

4.4.1 Το IPTC Συμβούλιο

Προς το στόχο της ανάπτυξης τεχνικών συγκεκριμένα για τις τηλεπικοινωνίες κινείται το IPTC (International Press Telecommunications Council[17]). Το IPTC είναι ένα συμβούλιο που αποτελείται από τα μεγαλύτερα ειδησεογραφικά προκτορεία και επιχειρήσεις, το οποίο αναπτύσσει και συντηρεί τεχνικά πρότυπα που στοχεύουν στην βελτίωση της ανταλλαγής ειδησεογραφικής πληροφορίας. Αυτή τη στιγμή τα μέλη του υπολογίζονται περίπου σε 55 εταιρείες και οργανισμούς από τον κόσμο των ειδήσεων.

Το συμβούλιο ιδρύθηκε το 1965 και κατά καιρούς ασχολήθηκε με διάφορες δραστηριότητες που αφορούσαν την ανάπτυξη προτύπων και τεχνικών που στοχεύουν πάντα στην προώθηση των υπηρεσιών τους. Τα τελευταία χρόνια με την έλευση της ηλεκτρονικής δημοσίευσης το IPTC προσανατολίζεται προς την ανάπτυξη προτύπων ειδησεογραφικού περιεχομένου και μεταδεδομένων, όπως τα NewsML, SportsML, ProgramGuideML και EventsML, NIFT (News Industry Text Format) και IIM (Information Interchange Model).

Η ανάγκη για την δημιουργία της NewsML (News Markup Language) απορρέει από την συνεχή αύξηση στην παραγωγή και χρήση των ειδήσεων σε όλο τον κόσμο με την ταυτόχρονη διάδοση του Διαδικτύου. Δημιουργήθηκε με σκοπό να αποτελέσει ένα πρότυπο αναπαράστασης και διαχείρισης ηλεκτρονικών ειδήσεων που απευθύνεται τόσο στους παραγωγούς όσο και στους τελικούς χρήστες. Επιπλέον, το πρότυπο σχεδιάστηκε έτσι ώστε να παρέχει ευελιξία και να επιτρέπει επεκτασιμότητα, που να ανταποκρίνεται σε διαφορετικές ανάγκες χρηστών.

Η NewsML έχει τη μορφή ενός XML εγγράφου (άλλωστε η δομή της βασίζεται στην XML), το οποίο αποτελείται από συστατικά και στοιχεία, τα οποία χρησιμοποιούνται για την δόμηση και την επεξεργασία της ειδησεογραφικής πληροφορίας. Γενικά, η NewsML είναι σχεδιασμένη έτσι

[17] www.iptc.org

ώστε να διατηρεί τα μεταδεδομένα όσο το δυνατον πιό κοντα στην εννοιολογική σημασία του αντικειμένου, ενώ παράλληλα πολλά από τα μεταδεδομένα είναι προαιρετικά.

Πέρα από τα σχήματα ανταλλαγής ειδησεογραφικού περιεχομένου το IPTC προώθησε και στη δημιουργία και συντήρηση ενός συνόλου θεμάτων (a set of topics), τα οποία έχουν την ιδιότητα να προσδίδουν αξία μεταδεδομένων (metadata values) σε αντικείμενα όπως κείμενα, φωτογραφίες, αρχεία ήχου και εικόνας. Κάτι τέτοιο επιτρέπει να επιτυγχάνεται μονιμότητα και συνέπεια στην ορολογία, καθώς κωδικοποιούνται τα ειδησεογραφικά μεταδεδομένα. Το σύνολο αυτό των θεμάτων αποτελεί τους IPTC NewsCodes, οι οποίοι ουσιαστικά δεν είναι τίποτε άλλο από ταξινομίσεις μεταδεδομένων για την ειδησεογραφία.

Οι IPTC NewsCodes για μεγαλύτερη ευκολία στη διαχείριση χωρίζονται σε 28 ξεχωριστά σύνολα (NewsCodes sets), τα οποία σχετίζονται μ' ένα συγκεκριμένο θέμα. Για παράδειγμα, το σύνολο "Audiocoders" περιέχει λεξιλόγιο για διάφορα είδη λογισμικού για ήχο. Επίσης, οι NewsCodes που περιέχονται σε κάθε σύνολο χρησιμοποιούνται ως συστατικά μεταδεδομένων (metadata elements) στις διάφορες φόρμες ανταλλαγής ειδήσεων. Τους News Codes μπορεί να τους βρεί κανείς στον ιστότοπο: (http://www.iptc.org/NewsCodes/). Διατίθενται και σε XML αρχεία.

Το παλαιότερο από τα σύνολα των NewsCodes είναι το Subject Reference System (SRS), που αναπτύχθηκε σε συνεργασία με την Newspaper Association of America[18]. Το SRS λειτουργεί ως ελεγχόμενο λεξιλόγιο για την κατηγοριοποίηση ειδησεογραφικών αντικειμένων και βασίζεται σε πρότυπα όπως το NewsML, το NIFT και το IIM και επίσης σε άλλα ταξινομικά συστήματα. Δεν περιέχεται στην λίστα των 28 NewsCodes, αλλά 5 από αυτούς αποτελούν τμήματά του (οι Genre, Media Type, Newsitem Type, Subject Code και Subject Qualifier).

[18] http://www.naa.org/

Ο Subject Code είναι ένα σύστημα περιγραφής περιεχομένου από ένα καλά ορισμένο σύνολο όρων. Οι όροι δίνονται με τη μορφή ιεραρχίας τριών επιπέδων. Τα θέματα του επιπέδου *Subject* περιγράφουν ειδησεογραφικό περιεχόμενο υψηλού επιπέδου, π.χ. *Arts, culture and entertainment.* Έπειτα οι όροι του *SubjectMatter* παρέχουν ακόμα μεγαλύτερη ανάλυση στην περιγραφή και αποτελούν υποκατηγορίες του προηγούμενου επιπέδου, π.χ. *music*. Τέλος, το *SubjectDetail* παρέχει ακόμα μεγαλύτερη ανάλυση και αντίστοιχα αποτελεί υποκατηγορία του *SubjectMatter,* π.χ. *classical music.*

Παρακάτω παρουσιάζονται οι βασικές κατηγορίες που περιέχονται στο επίπεδο *Subject*, με την ελληνική μετάφραση την οποία επιμελήθηκε το Μακεδονικό Πρακτορείο Ειδήσεων[19]. Τα μεταδεδομένα αυτά που παρέχει το IPTC χρησιμοποιήθηκαν για την κατηγοριοποίηση του ειδησεογραφικού περιεχομένου του παρόντος Συστήματος Εξατομικευμένης Ενημέρωσης και περιελαμβάνονται στην Βάση Περιεχομένου του Συστήματος. Οι 17 κύριες Θεματικές Κατηγορίες του IPTC εμφανίζονται στον πίνακα 1.

Arts, Culture & Entertainment (ACE)	Τέχνη, Πολιτισμός & Διασκέδαση
Crime, Law & Justice (CLJ)	Έγκλημα, Νομοθεσία και Δικαιοσύνη
Disasters & Accidents (DIS)	Καταστροφές & Ατυχήματα
Economy, Business & Finance (FIN)	Οικονομία, Επιχειρήσεις & Δημοσιονομία
Education (EDU)	Εκπαίδευση
Environmental issues (ENV)	Περιβαλλοντικά θέματα
Health (HTH)	Υγεία
Human Interest (HUM)	Ανθρωπιστικά ενδιαφέροντα
Labor (LAB)	Εργασία
Lifestyle & Leisure (LIF)	Τρόπος ζωής & Ελεύθερος χρόνος
Politics (POL)	Πολιτική
Religion & Belief (REL)	Θρησκεία & Πίστη
Science & Technology (SCI)	Επιστήμη και Τεχνολογία
Social Issues (SOI)	Κοινωνικά ζητήματα
Sport (SPO)	Αθλητισμός

[19] www.mpa.gr

Unrest, Conflicts & War (WAR)	Αναταραχές, διαμάχες & Πόλεμος
Weather (WEA)	Καιρός

Πίνακας 1: Βασικές θεματικές κατηγορίες του IPTC

Στις 17 αυτές βασικές κατηγορίες το Μακεδονικό Πρακτορείο πρόσθεσε επίσης μία επιπλέον κατηγορία που ανταποκρίνεται στις ανάγκες της Ελληνικής ειδησεογραφίας, την *Ομογένεια - Greek Living Abroad (GLA)*. Επιπλέον, για τις ανάγκες του Συστήματος στη Βάση Δεδομένων περιεχομένου (Content Indexing Database) προστέθηκαν 2 ακόμα κατηγορίες, η *Ελλάδα (Greece)* και τα *Διεθνή (International),* καθώς εμφανίζονται στην κατηγοριοποίηση των ειδησεογραφικών πηγών από τις οποίες το σύστημα εξάγει πληροφορίες. Από τις παραπάνω κατηγορίες του IPTC News Codes αυτές που εμφανίζονται στην κατηγοριοποίηση των ειδησεογραφικών πηγών και από τις οποίες το σύστημα εξάγει ειδήσεις είναι οι εξής: Πολιτισμός, Οικονομία, Πολιτική, Επιστήμη και Τεχνολογία, Κοινωνικά ζητήματα, Αθλητισμός, Καιρός, Διεθνή, Ελλάδα.

Η ιεραρχία του SRS χρησιμοποιήθηκε ως βάση για την ανάπτυξη της SRS οντολογίας, η οποία πραγματοποιήθηκε στα πλαίσια του Neptuno Project[20]. Υπεύθυνοι του προγράμματος είναι δύο πανεπιστήμια (το Universidad Autonoma de Madrid[21] και το Universitat de Lleida[22]), μία ειδησεογραφική εταιρεία (Diari SERGE) και ένας τεχνολογικός προωθητής (iSOCO, S.A.). Η οντολογία βασίζεται στα πρότυπα NewsML και SRS του IPTC σύμφωνα με τα οποία προχώρησε σε κάποιες επεκτάσεις. Η σύνταξη πραγματοποιήθηκε σε RDF έπειτα από μετάφραση της XML αναπαράστασης που παρέχεται στον ιστότοπο του IPTC. Η μετάφραση έγινε αυτόματα με τη βοήθεια ενός προγράμματος Java.

Την οντολογία SRS μπορεί να την κατεβάσει κανείς από την ιστοσελίδα: http://nets.ii.uam.es/neptuno/iptc/

[20] http://nets.ii.uam.es/neptuno
[21] http://www.uam.es/
[22] http://www.uam.es/

4.4.2 PRISM Initiative

Με την ανάπτυξη προτύπων για την ειδησεογραφία και γενικότερα για τον εκδοτικό τομέα ασχολείται, επίσης, η ομάδα εργασίας του PRISM[23] (Publishing Requirements for Industry Standard Metadata), η οποία συγκροτείται από εκδότες και πωλητές λογισμικού για ηλεκτρονικές εκδόσεις. Το PRISM ορίζει ένα προτυποποιημένο λεξιλόγιο XML μεταδεδομένων για την διαχείριση, συλλογή και επεξεργασία ειδήσεων, περιοδικών, βιβλίων, καταλόγων και γενικά δημοσιογραφικού περιεχόμενου. Χρησιμοποιεί συγκεκριμένα πρότυπα όπως η XML, RDF, το Dublin Core και διάφορες φόρμες του ISO για τοποθεσίες, ημερομηνίες, γλώσσες. Επιπλέον, το PRISM παρέχει ένα πλαίσιο εργασίας για την ανταλλαγή και διατήρηση περιεχομένου και μεταδεδομένων, την συλλογή συστατικών (elements) για την περιγραφή του περιεχομένου αυτού και τέλος ένα σύνολο ελεγχόμενων λεξιλογίων που θα ορίζει τις τιμές των συστατικών αυτών.

Τα μεταδεδομένα αντιστοιχούν σε υπερβολικά γενικές κατηγορίες πληροφοριών καλύπτοντας σχεδόν τα πάντα. Το πεδίο δράσης του PRISM καθορίζεται από τις ανάγκες των εκδοτών να λαμβάνουν, να εντοπίζουν και να διαθέτουν το περιεχόμενό τους. Κυρίως επικεντρώνεται στην απόδοση μεταδεδομένων για:

- Γενικού σκοπού περιγραφές των πηγών, οι οποίες μελετούνται ως ενιαίο σύνολο
- Ειδίκευση στις σχέσεις των πηγών μεταξύ τους
- Καθορισμό των πνευματικών δικαιωμάτων και αδειών
- Περιγραφή «εσωτερικών» (inline) μεταδεδομένων (markup within the resource itself)

Το PRISM συνιστά ένα σημαντικό βοήθημα για εφαρμογές Σημασιολογικού Ιστού στον ειδησεογραφικό χώρο. Παρόλ' αυτά όμως, δεν καλύπτει

[23] www.prismstandard.org

πρότυπα κωδικοποίησης ειδησεογραφικού περιεχομένου, σε αντίθεση με την NewsML, το SRS ή το NIFT. Επιπλέον, δεν προσφέρει πλούσια λογική εκφραστικότητα (όπως η OWL), απαραίτητη για τον εφοδιασμό ανεπτυγμένων τεχνολογιών έξυπνου περιεχομένου για το Σημασιολογικό Ιστό. Τέλος, δεν τίθεται το θέμα της πολυγλωσσικότητας, ένα θέμα ιδιάζουσας σημασίας για την ειδησεογραφική αγορά της Ευρώπης, η οποία συγκροτείται από πολλούς διαφορετικούς πολιτισμούς.

4.4.3 NEWS Project

Μία σημαντική προσπάθεια πραγματοποιείται από την IST (Information Society Technologies) μέσω του προγράμματος News Engine Web Services (NEWS Project[24]). Γενικότερος στόχος του προγράμματος είναι η αξιοποίηση των νέων τεχνολογιών για την ανάπτυξη εργαλείων που θα χρησιμοποιηθούν από ειδησεογραφικές υπηρεσίες. Οι τεχνολογίες αυτές βασίζονται στην ολοκλήρωση μοντέλων γνώσης ενός συγκεκριμένου πεδίου χρησιμοποιώντας ήδη υπάρχουσες οντολογίες υψηλού επιπέδου και πρότυπα που έχουν αναπτυχθεί στα πλαίσια του Σημασιολογικού Ιστού. Με τον τρόπο αυτό κατευθύνονται προς την ανάπτυξη ενός προτυποποιημένου λεξιλογίου για τον σημασιολογικό σχολιασμό αντικειμένων ειδησεογραφικής πληροφορίας. Με τον τρόπο αυτό ανοίγονται νέοι ορίζοντες για τους χρήστες που αφορούν την πρόσβαση, επιλογή και εξατομίκευση πολυμεσικού και πολυγλωσσικού ειδησεογραφικού περιεχομένου.

Στα πλαίσια και στους στόχους του προγράμματος εντάσσεται και η ανάπτυξη της NEWS Ontology. Η οντολογία αυτή σχεδιάζεται, τουλάχιστον σ' ένα πρώτο στάδιο, για να εξυπηρετήσει τις ανάγκες του συστήματος NEWS. Καθώς, λοιπόν, ακόμα το σύστημα βρίσκεται στο στάδιο κατασκευής υπάρχουν πολλά ζητήματα εκείνα που απασχολούν τους δημιουργούς του. Συνοπτικά αναφέρονται, η επιλογή της γλώσσας

[24] www.news-project.com

αναπαράστασης της οντολογίας, η ολοκλήρωση (integration) βάση ήδη υπαρχόντων συστημάτων, όπως το IPTC και το PRISM, εμβάθυνση στις υπηρεσίες που προσφέρουν οι οντολογίες στο συγκεκριμένο σύστημα, διατήρηση μίας συγκεκριμένης μεθοδολογίας στην πορεία της ανάπτυξης του συστήματος.

4.4.4 eBiquity News Ontology

Η ερευνητική ομάδα του UMBC (University of Maryland, Baltimore County) eBiquity έχει ως αντικείμενο έρευνας την αλληλεπίδραση μεταξύ των κινητών υπολογιστών (mobile computing), pervasive computing, multi-agent systems και τεχνητής νοημοσύνης, και web-based υπηρεσιών. Επίσης, τα ερευνητικά τους ενδιαφέροντα προσεγγίζουν περιοχές, όπως η διαχείριση γνώσης, η εξατομίκευση και η εξόρυξη δεδομένων (web/data mining).

Ένα από τα προγράμματα του UMBC eBiquity είναι και η ανάπτυξη οντολογιών. Αυτή τη στιγμή, στην ιστοσελίδα της eBiquity (http://ebiquity.umbc.edu/ontology/) διατίθενται οντολογίες για έντεκα διαφορετικά πεδία γνώσης, οι οποίες είναι βασισμένες στην OWL. Μία από αυτές είναι και η eBiquity News Ontology. Στην παρακάτω εικόνα παρουσιάζονται οι κλάσεις της eBiquity News ontology:

Classes
http://ebiquity.umbc.edu/ontology/news.owl#News
↳ label: News
↳ restriction:
↳ on property: http://ebiquity.umbc.edu/ontology/news.owl#title
↳ cardinality: 1
↳ restriction:
↳ on property: http://ebiquity.umbc.edu/ontology/news.owl#publishedOn
↳ max cardinality: 1
↳ restriction:
↳ on property: http://ebiquity.umbc.edu/ontology/news.owl#description
↳ max cardinality: 1
↳ restriction:
↳ on property: http://ebiquity.umbc.edu/ontology/news.owl#uri
↳ max cardinality: 1

Εικόνα 7: Κλάσεις της eBiquity News ontology σε σύνταξη OWL

Εκτός από τις οντολογίες που κυκλοφορούν ήδη στον Ιστό μπορεί κάποιος να αναπτύξει μία οντολογία ειδησεογραφικού (ή και άλλου περιεχομένου) βασιζόμενος σε ήδη υπάρχουσες οντολογίες ή να επαναχρησιμοποιήσει κάποιες άλλες (βλ. Κεφ. 4.2). Υπάρχουν βιβλιοθήκες οντολογιών, όπως η Ontolingua ontology library (http://www.ksl.stanford.edu/software/ontolingua/) ή η DAML ontology library (http://www.daml.org/ontologies).

Οι ανάγκες του Συστήματος Εξατομικευμένης Ενημέρωσης μας οδήγησαν τελικά στην επιλογή των IPTC News Codes για την κατηγοριοποίηση των ειδήσεων που προβάλει το σύστημα. Αυτό συμβαίνει διότι το σύστημα έχει ανάγκη από μία απλή κατηγοριοποίηση του περιεχομένου του, χωρίς την εφαρμογή μίας σύνθετης οντολογίας η οποία να βασίζεται σε κάποια προτυποποιημένη γλώσσα, όπως για παράδειγμα η OWL. Οι IPTC News Codes αποδείχθηκαν οι πλέον κατάλληλοι για την εφαρμογή αυτή και το μόνο που χρειάστηκε τελικά ήταν η προσθήκη δύο επιπλέον πεδίων (Ελλάδα και Διεθνή).

5. Εξαγωγή πληροφορίας

Στην ενότητα αυτή εξετάζεται η διαδικασία εξαγωγής πληροφορίας που εφαρμόστηκε στο παρόν σύστημα εξατομικευμένης ενημέρωσης από ειδικά προγράμματα (wrappers), καθώς επίσης και οι αλλαγές που πραγματοποιήθηκαν στη Βάση Περιεχομένου την οποία τροφοδοτούν οι wrappers αυτοί. Πριν όμως την περιγραφή του συστήματος, γίνεται μία εισαγωγή στην εξαγωγή πληροφορίας (Information Extraction).

5.1 Εισαγωγή στην εξαγωγή πληροφορίας

Ο μεγάλος όγκος δεδομένων που υπάρχουν στον Ιστό, έδωσαν αφορμή για έρευνα και μελέτη πάνω στην εξαγωγή χρήσιμων πληροφοριών από αυτά. Η εξαγωγή πληροφορίας ασχολείται με την εξαγωγή χρήσιμης πληροφορίας από διάφορες πηγές περιεχομένου. Για παράδειγμα, μία τυπική εργασία εξαγωγής πληροφορίας θα μπορούσε να είναι ο εντοπισμός των "τρομοκρατικών επιθέσεων" που αναφέρονται στις εφημερίδες. Σ' αυτό το σημείο πρέπει να γίνει μία επισήμανση προκειμένου να μην υφίσταται σύγχυση μεταξύ της Εξαγωγής πληροφορίας και της Ανάκτησης Πληροφορίας (Information Retrieval). Η τελευταία ασχολείται με την συγκέντρωση κειμένων (π.χ. ιστοσελίδων) σχετικών με ένα πεδίο, ενώ η Εξαγωγή πληροφορίας με την εξαγωγή σχετικής πληροφορίας από τις σελίδες.

Τα συστήματα αυτά που συλλέγουν και επεξεργάζονται δεδομένα από πολλαπλές πηγές του Ιστού ονομάζονται *Web Information Integration Systems* (Συστήματα Ενοποίησης Πληροφορίας) και το ενδιαφέρον για την ανάπτυξή τους διαρκώς μεγαλώνει. Το μειονέκτημα, όμως, είναι πως τα έγγραφα που δημοσιεύονται στον Ιστό αρχικά σχεδιάστηκαν για χρήση από τους ανθρώπους, με αποτέλεσμα η αυτοματοποίηση της διαδικασίας εξαγωγής πληροφορίας να αποτελεί μία αρκετά δύσκολη εργασία.

Τα *Web Information Integration Systems* χρησιμοποιούν ένα σύνολο προγραμμάτων εξαγωγής πληροφορίας, τους λεγόμενους wrappers, και ένα ενδιάμεσο πληροφορίας (information mediator), ο οποίος αποτελεί τον μεσάζοντα ανάμεσα στα προγράμματα εξαγωγής πληροφορίας και τους χρήστες. Στόχος του είναι να ενοποιεί τις πληροφορίες σε μία πιο ολοκληρωμένη μορφή. Μία τυπική εφαρμογή wrapper χρησιμοποιείται για να εξάγει δεδομένα από δομημένες ιστοσελίδες (σε HTML μορφή) συμπληρώνοντας στη συνέχεια μία σχεσιακή βάση δεδομένων. Τα προγράμματα wrappers εξάγουν την πληροφορία εκμεταλλευόμενα τη δομή των ιστοσελίδων, δίχως τη χρήση εκτενούς γλωσσικής επεξεργασίας, επιτρέποντας την αναζήτηση πληροφοριών με δομημένες επερωτήσεις, καθώς και την παρουσίασή τους σε συνοπτική, ενοποιημένη μορφή.

Η διαρκής εξέλιξη του Ιστού και των νέων τεχνολογιών οδηγεί σε ένα ασταθές περιβάλλον δυσκολεύοντας την απόδοση των wrappers. Καταρχήν, ο αριθμός των πηγών αυξάνεται καθιστώντας δύσκολη την αναζήτηση και εξαγωγή πληροφοριών απο τους wrappers. Το σημαντικότερο μειονέκτημα, όμως, είναι ότι οι wrappers αχρηστεύονται όταν οι σελίδες, από τις οποίες εξάγουν πληροφορίες αλλάζουν δομή, ένα φαινόμενο αρκετά συχνό στο περιβάλλον του Παγκόσμιου Ιστού. Ένα πρόβλημα, το οποίο αντιμετωπίστηκε και στο Σύστημα Εξατομικευμένης Ενημέρωσης, όταν το CNN, μία από τις πηγές περιεχομένου του συστήματος, άλλαξε τη δομή του.

Οι wrappers χρησιμοποιήθηκαν από το Σύστημα Εξατομικευμένης Ενημέρωσης για την εξαγωγή των ειδήσεων από συγκεκριμένες ειδησεογραφικές πηγές του Παγκόσμιου Ιστού. Στα πλαίσια της παρούσας εργασίας έγινε έρευνα για την επιλογή των πηγών αυτών και για τις κανονικές εκφράσεις που χρησιμοποιήθηκαν από τους wrappers για την εξαγωγή των ειδήσεων.

5.2 Ειδησεογραφικές πηγές απ' όπου εξάγει πληροφορίες το Σύστημα Εξατομικευμένης Ενημέρωσης

Ζητούμενο για το παρόν σύστημα ενημέρωσης είναι να αποτελείται από πηγές γενικού κυρίως ενδιαφέροντος, τόσο ελληνικής όσο και ξένης ειδησεογραφίας. Αρχικά, οι wrappers αντλούσαν πληροφορίες από τις ιστοσελίδες του CNN[25] και της Ναυτεμπορικής[26]. Στόχος της παρούσας πτυχιακής εργασίας ήταν να προστεθούν στο σύστημα νέες πηγές ειδησεογραφικού περιεχομένου. Στην πορεία, όμως, παρουσιάστηκε πρόβλημα με το CNN, το οποίο αλλάζει τόσο τακτικά τη δομή των ιστοσελίδων του, ώστε να καθίσταται αδύνατη η εξαγωγή σωστών πληροφοριών από τους wrappers. Το γεγονός αυτό οδήγησε στην αναγκαστική απόρριψη του CNN ως πηγή περιεχομένου του συστήματος. Οι πηγές ειδησεογραφικού περιεχομένου που συμπεριλήφθηκαν στο σύστημα τελικά, εκτός από την Ναυτεμπορική είναι η Ελευθεροτυπία[27], ο Antenna[28], το bbc.greek[29], η Le monde[30] και το yahoo[31].

Σε ένα πρώτο στάδιο η επιλογή των ειδησεογραφικών πηγών έγινε με βάση το περιεχόμενο και την επισκεψιμότητα των πηγών. Στη συνέχεια παρουσιάστηκαν και κάποιοι άλλοι παράγοντες, οι οποίοι αναπόφευκτα επηρέασαν και αυτοί με τη σειρά τους τα κριτήρια επιλογής. Οι παράγοντες αυτοί ήταν τεχνικού κυρίως χαρακτήρα, που αφορούσαν τη δομή της URL διεύθυνσης ή του κώδικα της ιστοσελίδας της πηγής και που εμπόδιζαν με την πολυπλοκότητά τους την λειτουργία των wrappers. Γι' αυτό το λόγο, για παράδειγμα στάθηκε αδύνατο να προσθέσουμε στις πηγές μας και την Καθημερινή, καθώς στη σύνταξη των URL των σελίδων της συμπεριλαμβάνεται η ημερομηνία και ένας άλλος μεταβλητός αριθμός, προσδίδοντας στις URL έναν ασταθή χαρακτήρα.

[25] http://www.cnn.com/
[26] http://www.naftemporiki.gr/
[27] http://www.enet.gr/online/online
[28] http://news.antenna.gr/
[29] http://www.bbc.co.uk/greek/
[30] www.lemonde.fr
[31] http://news.yahoo.com

Από τις πρώτες επιλογές ήταν η Ελευθεροτυπία, εφημερίδα αρκετά καταξιωμένη στον Ελλαδικό χώρο, που εκδίδεται τόσο σε ηλεκτρονική όσο και σε έντυπη μορφή. Η ηλεκτρονική έκδοση της Ελευθεροτυπίας διατίθεται δωρεάν στον Ιστό, όπως όλες άλλωστε οι ειδησεογραφικές πηγές που περιλήφθηκαν στο Σύστημα. Από τις θεματικές κατηγορίες, με τις οποίες ταξινομείται το υλικό της, στο σύστημά μας συμπεριλάβαμε τις: *πολιτική, Ελλάδα, οικονομία, τέχνες, αθλητισμός, κόσμος (Διεθνή)*. Μία ακόμη εφημερίδα που εκδίδεται τόσο σε ηλεκτρονική όσο και σε έντυπη μορφή είναι η Γαλλική Le Monde. Η Θεματική κατηγοριοποίηση της Le Monde είναι αρκετά διευρυμένη παρέχοντας στον αναγνώστη μεγαλύτερη ευκολία στην αναζήτηση των ειδήσεων. Από όλες τις κατηγορίες τελικά επιλέχθηκαν οι εξής: *International* (Διεθνή), *societe* (Κοινωνικά ζητήματα), *sports* (Αθλητικά), *sciences* (Επιστήμη), *culture* (Πολιτισμός).

Εκτός από εφημερίδες, η επιλογή για ειδησεογραφικές πηγές έγινε, επίσης, ανάμεσα από ηλεκτρονικά τηλεοπτικά μέσα που προωθούν ειδησεογραφικό υλικό. Οι τηλεοπτικές πηγές που επιλέχθηκαν είναι ο Antenna και το bbc. Οι ειδήσεις του Antenna στο Διαδίκτυο διανέμονται μέσω του ιστότοπου http://news.antenna.gr/ και οι κατηγορίες από τις οποίες αντλούνται ειδήσεις για το Σύστημα Εξατομικευμένης Ενημέρωσης είναι οι εξής: *πολιτική, κοινωνία, διεθνή, οικονομία, αθλητικά, πολιτισμός, επιστήμη, καιρός.* Για το bbc από την άλλη, οι ειδήσεις εξάγονται από την ελληνική έκδοση, bbc.greek. Συγκεκριμένα, οι ειδήσεις εξάγονται από τις θεματικές κατηγορίες του bbc.greek *διεθνή, Ελλάδα – Κύπρος, οικονομία, πολιτισμός, αθλητικά.*

Τέλος, μία ακόμη πηγή που συμπεριλήφθηκε στο σύστημα αποτελεί και το yahoo.news (http://rss.news.yahoo.com/rss/). Η επιλογή αυτή πραγματοποιήθηκε στα πλαίσια της εργασίας των φοιτητών Αλέξανδρου Μουζακίδη και Χρήστου Ντούτση και αφορά την μελέτη των RSS Feeds, με σκοπό την εφαρμογή τους στο Σύστημα Εξατομικευμένης Ενημέρωσης. Τα RSS σε αντίθεση με τους wrappers, τους οποίους χρησιμοποιεί το σύστημα για να εξάγει πληροφορίες από τις υπόλοιπες πηγές, παρέχουν μεγαλύτερη ευκολία και εξυπηρέτηση στο σύστημα.

Οι προαναφερθέντες πηγές, εκτός του yahoo.news, επιλέχθηκαν για το παρόν σύστημα στα πλαίσια της παρούσας εργασίας. Το σύστημα, όμως, όπως ήδη αναφέρθηκε, αντλεί ειδήσεις και από την Ναυτεμπορική, συγκεκριμένα από τις θεματικές κατηγορίες του ιστότοπου, *οικονομία, αθλητικά, πολιτικά, κόσμος, πολιτιστικά*.

Η εξαγωγή των ειδήσεων από τις παραπάνω πηγές πραγματοποιείται από προγράμματα εξαγωγής πληροφορίας (wrappers). Η διαδικασία που ακολουθήθηκε για την υλοποίηση τους παρουσιάζεται παρακάτω, αφού πρώτα γίνεται μία εισαγωγή στον τρόπο λειτουργίας τους για το παρόν σύστημα εξατομικευμένης ενημέρωσης.

5.3 Προγράμματα εξαγωγής πληροφορίας (wrappers) και η χρήση τους στο Σύστημα Εξατομικευμένης Ενημέρωσης

Το παρόν σύστημα εξατομικευμένης ενημέρωσης εξάγει αυτόματα πληροφορίες από τις παραπάνω ειδησεογραφικές πηγές του Παγκόσμιου Ιστού και στη συνέχεια τις κάνει διαθέσιμες στους χρήστες ανάλογα με τα ενδιαφέροντα τους. Όπως αναφέρθηκε και στο κεφάλαιο 3 η εξαγωγή των ειδήσεων είναι μία εργασία που επιτελείται από τον Content Scanner. Ο Content Scanner είναι μία αυτόνομη εφαρμογή που λειτουργεί ανεξάρτητα από τις εξατομικευμένες υπηρεσίες του συστήματος. Κατά τακτά χρονικά διαστήματα ενημερώνει τη Βάση Δεδομένων του συστήματος με τις καινούριες ειδήσεις που υπάρχουν στις πηγές. Ως είσοδο, παίρνει ορισμένες παραμέτρους από τη Βάση Περιεχομένου (την URL διεύθυνση και τις κανονικές εκφράσεις) και ως έξοδο επιστρέφει τις νέες ειδήσεις στη Βάση.

Όπως φαίνεται από την εικόνα 3 ο Content Scanner αποτελείται από wrappers που εξάγουν την επιθυμητή είδηση από την ιστοσελίδα – στόχο. Οι wrappers αυτοί δεν δημιουργούνται αυτόματα αλλά ορίζονται σε σχέση με κάποια διεύθυνση στον Ιστό (URL) και με βάση ορισμένες κανονικές

εκφράσεις (regular expressions). Το παρόν σύστημα διαθέτει τον wrapper επιπέδου 1 και τον wrapper επιπέδου 2. Ο wrapper 1 αντλεί τις διευθύνσεις των ειδήσεων και τις στέλνει στο wrapper 2, ο οποίος με τη σειρά του αντλεί τον τίτλο, την πρώτη σειρά του άρθρου και ολόκληρο το άρθρο.

Συγκεκριμένα, και για να γίνει κατανοητός ο τρόπος με τον οποίο εξάγονται οι ειδήσεις, χρειάζεται να αναφέρουμε ότι ο Content Scanner ακολουθεί την δομή της κάθε ιστοσελίδας ενημέρωσης. Έτσι, λοιπόν, ο Wrapper επιπέδου 1 παίρνει ως είσοδο από την Βάση Περιεχομένου την διεύθυνση των ιστοσελίδων (π.χ. *οικονομικά* της Ελευθεροτυπίας) και τις κανονικές εκφράσεις που θα χρησιμοποιηθούν στην εξαγωγή των οικονομικών ειδήσεων της Ελευθεροτυπίας κατόπιν με μία HTTP αίτηση ανακτά το περιεχόμενο αυτής της ιστοσελίδας. Χρησιμοποιώντας τις κανονικές εκφράσεις επεξεργάζεται την ιστοσελίδα και ανακτά τις διευθύνσεις όλων των οικονομικών άρθρων της Ελευθεροτυπίας.

Στη συνέχεια, ο wrapper επιπέδου 2 παίρνει ως είσοδο τις διευθύνσεις που έχει εξάγει ο wrapper 1 και τις κανονικές εκφράσεις από τη Βάση Περιεχομένου. Με μία HTTP αίτηση ανακτά το περιεχόμενο των διευθύνσεων αυτών και εξάγει τον τίτλο, την πρώτη πρόταση της είδησης και ολόκληρο το άρθρο σαν απλό κείμενο απαλλαγμένο από HTML tags. Η διεύθυνση, ο τίτλος και η πρώτη πρόταση του άρθρου αποθηκεύονται στη Βάση Περιεχομένου με σκοπό να χρησιμοποιηθούν από τον Content Selector κατά τη διάρκεια της εξατομίκευσης. Από σεβασμό προς τα δικαιώματα των πηγών δημοσίευσης των άρθρων, δεν παρουσιάζεται ολόκληρο το άρθρο από τον Content Presenter αλλά ένας υπερσύνδεσμος που οδηγεί στην ιστοσελίδα της πηγής.

Για να γίνει περισσότερο αντιληπτός ο τρόπος με τον οποίο δουλεύει ο content Scanner δίνεται το παρακάτω παράδειγμα. Σε ένα πρώτο στάδιο, ο wrapper 1 από τον πίνακα *source_category* της Βάσης Δεδομένων πληροφορείται τις διευθύνσεις των πηγών (πχ. http://www.naftemporiki.gr/news/static/wld.htm: URL της κατηγορίας

Διεθνή της Ναυτεμπορικής) στις οποίες θα ανατρέξει για να εξάγει τις διευθύνσεις των άρθρων. Έχοντας εντοπίσει πλέον τις διευθύνσεις των άρθρων τις "στέλνει" στον wrapper επιπέδου 2 για την εξαγωγή του τίτλου, της πρώτης σειράς και ολόκληρου το άρθρου. Για να εξαχθεί, για παράδειγμα, ο τίτλος μίας είδησης, ο wrapper 2 χρησιμοποιεί από τη Βάση Περιεχομένου της κανονικές εκφράσεις, που βρίσκονται στον πίνακα *content_wrapper* και συγκεκριμένα είναι τα πεδία "startOfTitleRegEx" και "endOfTitleRegEx". Στη Ναυτεμποτική οι κανονικές εκφράσεις που έχουν δοθεί είναι τα HTML tags <TITLE> και </TITLE> αντίστοιχα. Ως έξοδο, λοιπόν, ο scanner επιστρέφει στην Βάση Δεδομένων (στον πίνακα *news content*) όλες εκείνες τις πληροφορίες που άντλησε. Παρόμοια γίνεται η εξαγωγή όλων των πληροφοριών από τις ειδήσεις.

Ένα πρόβλημα του Content Scanner είναι πως οι wrappers δεν δημιουργούνται αυτόματα, οπότε όταν αλλάζει ο κώδικας που παράγει τις ιστοσελίδες των πηγών το σύστημα δεν εξάγει σωστές πληροφορίες, πρόβλημα το οποίο αναφέρθηκε και προηγουμένως σχετικά με το CNN.

Η εξαγωγή ειδήσεων από τις νέες ειδησεογραφικές πηγές επέβαλε την προσθήκη νέων κανονικών εκφράσεων στη Βάση Περιεχομένου, όπως επίσης και τη δημιουργία τελικά μίας νέας Βάσης. Η διαδικασία αυτή αποτέλεσε και την βασική εργασία της πτυχιακής αυτής και διεξάχθηκε στο Εργαστήριο Τεχνολογιών Γνώσεων και Λογισμικού του Ε.Κ.Ε.Φ.Ε Δημόκριτος. Προτού προχωρήσουμε στην παρουσίαση της διαδικασίας υλοποίησης των wrappers, παρουσιάζεται η αρχική Βάση Περιεχομένου, στην οποία βασιστήκαμε σε ένα πρώτο στάδιο για την μελέτη των κανονικών εκφράσεων και την εξαγωγή ειδήσεων. Η νέα Βάση Περιεχομένου παροτίθεται στο τελειυταίο κεφάλαιο (κεφάλαιο 5.6).

5.4 Αρχική Βάση Περιεχομένου

Σε ένα πρώτο στάδιο, όπως θα δούμε και παρακάτω, χρησιμοποιήθηκε η Βάση Περιεχομένου, όπως την είχε στήσει ο Αλεξόπουλος Άγγελος και στην πορεία δημιουργήθηκε η ανάγκη για την δημιουργία μίας νέας Βάσης Περιεχομένου. Πρωτού, λοιπόν, προχωρήσουμε στην παρουσίαση της έρευνας για τον καθορισμό των κονονικών εκφράσεων και της υλοποίησης των wrappers, θεωρούμε αναγκαία την παρουσίαση της πρωταρχικής Βάσης Περιεχομένου, ώστε να διευκολύνουμε τον αναγνώστη να έχει ολοκληρωμένη εικόνα της έρευνας αυτής.

Η Βάση Περιεχομένου αποτελείται από έξι πίνακες οι οποίοι περιέχουν όλη την πληροφορία για τις κατηγορίες των ειδήσεων, τις πηγές των ειδήσεων και τις κανονικές εκφράσεις που χρησιμοποιούν τα προγράμματα των wrappers.

- **Πίνακας news_category**

Στον πίνακα αυτό φυλάσσονται οι κατηγορίες ειδήσεων που εξάγονται από τον Content Scanner.

news _category	Περιγραφή
News_id	Το κλειδί του πίνακα
News_Subject	Η κατηγορία ειδήσεων (π.χ. Οικονομία, Αθλητισμός,…)

- **Πίνακας source_name**

Στον πίνακα αυτό φυλάσσονται οι ειδησεογραφικές πηγές από τις οποίες εξάγονται ειδήσεις από τον Content Scanner.

source_name	Περιγραφή
Source_id	Το κλειδί του πίνακα
Source_name	Το όνομα της πηγής (π.χ. Ναυτεμπορική)
Source_url	Η διεύθυνση στο Web (URL) της πηγής (π.χ.

	http://www.naftemporiki.gr)

- **Πίνακας source_category**

Στον πίνακα αυτό φυλάσσονται οι διευθύνσεις στο Web (URL) των ιστοσελίδων των πηγών για τις συγκεκριμένες κατηγορίες ειδήσεων από τις οποίες εξάγονται ειδήσεις από τον Content Scanner.

source_name	Περιγραφή
Id	Το κλειδί του πίνακα
url	Η διεύθυνση στο Web (URL) της πηγής για την συγκεκριμένη κατηγορία (π.χ. http://www.naftemporiki.gr/news/static/fin.htm)
category_id	Το κλειδί από τον πίνακα news_category, το οποίο αντιστοιχεί στην συγκεκριμένη κατηγορία
Source_id	Το κλειδί από τον πίνακα source_name, το οποίο αντιστοιχεί στην συγκεκριμένη πηγή

- **Πίνακας news_content**

Στον πίνακα αυτό φυλάσσονται οι ειδήσεις όπως εξάγονται από τον Content Scanner

news_content	Περιγραφή
News_id	Το κλειδί του πίνακα
Title	Ο τίτλος της είδησης
Category	Το κλειδί της κατηγορίας του πίνακα news_category, στην οποία ανήκει η είδηση
Ndate	Η ημερομηνία εξαγωγής της είδησης
Source	Το κλειδί της πηγής του πίνακα source_name, στην οποία ανήκει η είδηση
File	Το όνομα του αρχείου στο οποίο αποθηκεύεται τοπικά το άρθρο της είδησης
url	Η διεύθυνση στο Web (URL) της είδησης

Sentence	Η πρώτη φράση της είδησης

- **Πίνακας wrapper1**

Στον πίνακα αυτό φυλάσσονται οι πληροφορίες εισόδου του Content Scanner και συγκεκριμένα οι πληροφορίες που χρησιμοποιούν οι wrappers επιπέδου 1 για να εξάγουν τις διευθύνσεις των ειδήσεων που θα επεξεργασθούν οι wrappers επιπέδου 2.

wrapper1	Περιγραφή
Wrapp1id	Το κλειδί του πίνακα
SiteId	Το κλειδί του πίνακα source_category
Begreg	Κανονική έκφραση που δείχνει από ποιο σημείο να αρχίσει η εξαγωγή πληροφορίας (δηλαδή του URL)
Endreg	Κανονική έκφραση που δείχνει σε ποιο σημείο να σταματήσει η εξαγωγή πληροφορίας (δηλαδή του URL)
flagSentence	Κανονική έκφραση που δείχνει πότε να ξεκινήσει η εξαγωγή πληροφορίας. Όλα τα δεδομένα της ιστοσελίδας πριν από αυτή την κανονική έκφραση αγνοούνται.
NotMatch	Κανονική έκφραση που δείχνει ποια δεδομένα πρέπει να αγνοηθούν, όταν έχει αρχίσει ήδη η εξαγωγή πληροφορίας
Flag1	Παίρνει την τιμή 0 όταν υπάρχει flagSentence. Όταν δεν υπάρχει κάποια flagSentence και έχει τιμή ανύπαρκτη όπως NOTHING παίρνει τιμή 1.
Flag2	Αρχικοποιείται με την τιμή 1 όταν επιθυμούμε να εξαχθεί ένας συγκεκριμένος αριθμός διευθύνσεων. Όταν έχει τιμή 0 εξάγονται όλες οι διευθύνσεις των ειδήσεων και δεν λαμβάνεται υπ’ όψη η παράμετρος numberIterations.
NumberIterations	Πλήθος διευθύνσεων (μέγιστο) που επιθυμούμε να εξαχθούν Όταν θέλουμε να εξαχθούν όλες οι ειδήσεις από το συγκεκριμένο URL βάζουμε μια μεγάλη τιμή π.χ. 100

- **Πίνακας wrapper2**

Στον πίνακα αυτό φυλάσσονται οι πληροφορίες εισόδου του Content Scanner και συγκεκριμένα οι πληροφορίες που χρησιμοποιούν οι wrappers επιπέδου 2 για να εξάγουν τις πληροφορίες που αποθηκεύονται στον πίνακα news_content.

wrapper2	Περιγραφή
wrapp2id	Το κλειδί του πίνακα
wrapp1	Το κλειδί του πίνακα wrapper1
begregTitle	Κανονική έκφραση που δείχνει από ποιο σημείο να αρχίσει η εξαγωγή του τίτλου της είδησης
endregTitle	Κανονική έκφραση που δείχνει σε ποιο σημείο να σταματήσει η εξαγωγή του τίτλου της είδησης
flagSentence	Κανονική έκφραση που δείχνει πότε να ξεκινήσει η εξαγωγή της πρώτης φράσης της συγκεκριμένης είδησης. Όλα τα δεδομένα της ιστοσελίδας πριν από αυτή την κανονική έκφραση αγνοούνται.
BegregSentence	Κανονική έκφραση που δείχνει από που να ξεκινήσει να εξάγεται η πρώτη φράση της είδησης
begregSentence1	Εναλλακτική κανονική έκφραση που δείχνει από που να ξεκινήσει να εξάγεται η πρώτη φράση της είδησης. Χρησιμοποιείται όταν αλλάζει εντελώς ο τρόπος που ξεκινάει ένα άρθρο (π.χ. σε μερικές περιπτώσεις υπάρχει κάποια φωτογραφία ή κάποιος πίνακας)
EndregSentence	Κανονική έκφραση που δείχνει που να σταματήσει να εξάγεται η πρώτη φράση της είδησης
Flag	Αρχικοποιείται με την τιμή 0 που δείχνει ότι η flagSentence δεν έχει βρεθεί ακόμα στο HTML κείμενο.
begregFile	Κανονική έκφραση που δείχνει από που να ξεκινήσει να εξάγεται το άρθρο της είδησης
endregFile	Κανονική έκφραση που δείχνει που να σταματήσει να εξάγεται το άρθρο της είδησης
result	Boolean μεταβλητή η οποία είναι αληθής όταν το κυρίως άρθρο περιλαμβάνει την «πρώτη πρόταση», ενώ είναι

	ψευδής όταν δεν περιλαμβάνεται. Για παράδειγμα, τα άρθρα της Ναυτεμπορικής παίρνουν αληθή τιμή ενώ τα άρθρα του CNN παίρνουν ψευδή τιμή.

Στη παρακάτω ενότητα, όπου παρουσιάζεται η διαδικασία υλοποίησης των wrappers, ο αναγνώστης θα παρατηρήσει δύο διαφορετικούς τύπους γραφής των πεδίων Βάσης Περιεχομένου. Για λόγους πρακτικότητας τα πεδία της αρχικής Βάσης γράφονται με **bold** ενώ τα πεδία της νέας είναι γραμμένα σε *Comics Sans MS*.

5.5 Υλοποίηση των wrappers

Στην ενότητα αυτή παρουσιάζονται οι διαδικασίες που χρειάστηκε να πραγματοποιηθούν προκειμένου να μπορεί ο Content Scanner να αντλεί πληροφορίες από τις νέες ειδησεογραφικές πηγές, που επιλέχθηκαν για το Σύστημα Εξατομικευμένης Ενημέρωσης. Όπως αναφέρθηκε και στην προηγούμενη ενότητα, οι wrappers για την εξαγωγή πληροφορίας χρησιμοποιούν ορισμένες κανονικές εκφράσεις (regular expressions). Όπως είναι αναμενόμενο, για κάθε πηγή οι κανονικές εκφράσεις είναι διαφορετικές. Ο καθορισμός τους στηρίχθηκε στον HTML κώδικα της κάθε ιστοσελίδας, από την οποία θέλουμε να αντλήσουμε τις πληροφορίες που μας ενδιαφέρουν.

Η αναζήτηση των κανονικών εκφράσεων υπήρξε μία ιδιαίτερα χρονοβόρα διαδικασία, η οποία εγκυμονούσε διάφορα προβλήματα. Αρκεί, άλλωστε, να αναλογιστεί κανείς το πόσο διαφορετικές, από άποψη δομής, πηγές περιλήφθηκαν στο Σύστημα Εξατομικευμένης Ενημέρωσης. Σημαντικός οδηγός στην προσπάθεια να δοθούν όσο το δυνατόν αποτελεσματικότερες κανονικές εκφράσεις αποτέλεσαν τόσο τα Special JavaScript Regular Expressions[32] όσο, επίσης, και η εφαρμογή espresso. Τα Special JavaScript Regular Expressions παρέχουν κανόνες σύνταξης κανονικών εκφράσεων και αποτελούν ένα πρότυπο ευρέως διαδεδομένο. Από την άλλη η

[32] www.javascriptkit.com/javatutors/redev2.shtml

εφαρμογή espresso[33] χρησιμοποιήθηκε για ευκολία στην δοκιμή των κανονικών εκφράσεων πάνω στον κώδικα της κάθε ιστοσελίδας. Ο κώδικας μεταφέρονταν στην εφαρμογή (με ένα απλό copy-paste), όπου και δουλεύονταν οι κανονικές εκφράσεις, ώστε να εξασφαλίζεται η εγκυρότητά τους.

Προκειμένου να δοθεί μία σαφέστερη εικόνα για τον τρόπο με τον οποίο πραγματοποιήθηκε ο καθορισμός των κανονικών εκφράσεων χρησιμοποιούμε ένα παράδειγμα. Το εν λόγω παράδειγμα εστιάζεται στον καθορισμό κανονικής έκφρασης, με την οποία ο wrapper επιπέδου 1 ανασύρει τις διευθύνσεις των αθλητικών άρθρων από την Ναυτεμπορική. Σε ένα πρώτο στάδιο, έγινε αντιγραφή του κώδικα HTML της αρχικής σελίδας της κατηγορίας "*Αθλητισμός*" (http://www.naftemporiki.gr/news/static/spo.htm) στο espresso (ο HTML κώδικας φαίνεται στην **εικόνα 8**). Μέσα στον κώδικα εντοπίζεται η πληροφορία που θέλουμε να εξάγουμε (στην συγκεκριμένη περίπτωση οι διευθύνσεις των αθλητικών άρθρων), η οποία βρίσκεται ανάμεσα από ορισμένα tags. Τα tags αυτά είναι που χρησιμοποιήθηκαν για τον ορισμό των κανονικών εκφράσεων. Συγκεκριμένα, παρατηρούμε ότι οι διευθύνσεις των άρθρων τοποθετούνται ακριβώς πρίν το tag <a href=". Αν, όμως, ορίσουμε το συγκεκριμένο tag μόνο του ως κανονική έκφραση οι wrappers θα εξάγουν και άλλες άχρηστες πληροφορίες. Επομένως, η κανονική έκφραση πρέπει να διακρίνεται από το στοιχείο της μοναδικότητας μες το HTML κείμενο.

Στην **εικόνα 8** παρουσιάζεται μέρος του κώδικα της ιστοσελίδας:
http://www.naftemporiki.gr/news/static/spo.htm.
Με **bold** σημειώνονται οι κανονικές εκφράσεις που χρησιμοποιήθηκαν στο πεδίο "startOfLinkRegEx" του πίνακα *html_wrapper* για την εξαγωγή των διευθύνσεων των άρθρων, ενώ με *italics* οι ίδιες οι διευθύνσεις. Με τη βοήθεια της εφαρμογής espresso μπορεί να βρεθεί η κανονική έκφραση που εξάγει ακριβώς την πληροφορία που ενδιαφέρει και τίποτε

[33] www.ultrapico.com/Expresso.htm

περισσότερο. Στην προκειμένη περίπτωση η κανονική έκφραση: ‹font class="contenttextblack"><A href=" δίνει ως αποτέλεσμα τις παρακάτω διευθύνσεις:

1. http://www.naftemporiki.gr/news/static/05/09/28/1101843.htm"
2. http://www.naftemporiki.gr/news/static/05/09/28/1101779.htm
3. http://www.naftemporiki.gr/news/static/05/09/28/1101774.htm

```
<tr>
<td width=35 valign="top"><font class="contenttextlite"><b>28/9</b></font></td>
<td width=35 valign="top"><font class="contenttextlite"><b>21:08</b></font></td>
<td width=340 valign="top"><font class="contenttextblack"><A
href="http://www.naftemporiki.gr/news/static/05/09/28/1101843.htm"  class="menu"
title="" TARGET="_top" >Η 3η κλήρωση του SUPER 3</a></font></a></td>
</tr>
<tr>
<td width=35 valign="top"><font class="contenttextlite"><b>28/9</b></font></td>
<td width=35 valign="top"><font class="contenttextlite"><b>20:10</b></font></td>
<td width=340 valign="top"><font class="contenttextblack"><A
href="http://www.naftemporiki.gr/news/static/05/09/28/1101779.htm"  class="menu"
title="" TARGET="_top" >Στη διάθεση του Κόσμι ο Ιακουΐντα</a></font></a></td>
</tr>
<tr>
<td width=35 valign="top"><font class="contenttextlite"><b>28/9</b></font></td>
<td width=35 valign="top"><font class="contenttextlite"><b>20:07</b></font></td>
<td width=340 valign="top"><font class="contenttextblack"><A
href="http://www.naftemporiki.gr/news/static/05/09/28/1101774.htm"  class="menu"
title="" TARGET="_top" >Με Πιτζόνια η Γουίλιαμς στα γκραν πρι Ιαπωνίας και
Κίνας</a></font></a></td></tr>
```

Εικόνα 8: Κώδικας της ιστοσελίδας:
http://www.naftemporiki.gr/news/static/spo.htm.

Παρόμοια με την παραπάνω διαδικασία βρέθηκαν όλες οι κανονικές εκφράσεις, που χρησιμοποιήθηκαν από τους wrappers για την εξαγωγή πληροφοριών από τις πηγές. Κατά τη διάρκεια της αναζήτησης των κατάλληλων κανονικών εκφράσεων απαντήθηκαν διάφορες δυσκολίες, οι οποίες ποικίλουν ανάλογα με τη δομή της κάθε πηγής.

Η Ναυτεμπορική, η Ελευθεροτυπία και το bbc.greek παρουσιάζουν δομή, που ευνοεί τον καθορισμό και την εφαρμογή των wrappers, μια και οι

πληροφορίες που ζητάμε να εξάγουμε βρίσκονται πάντα ανάμεσα σε μοναδικά strings μέσα στον κώδικα (όπως στον κώδικα της εικόνας 8). Από την άλλη όμως, στον Antenna και τη Le Monde, η ίδια αυτή αποστολή αποδεικνύεται σαφώς πιό πολύπλοκη, καθώς, πολύ συχνά τα tags που προηγούνται της ζητούμενης πληροφορίας ποικίλουν. Με τον τρόπο αυτό ο scanner αδυνατεί να εξάγει την κατάλληλη πληροφορία. Στις περιπτώσεις αυτές, το πρόβλημα αντιμετωπίστηκε εφαρμόζοντας τα Special JavaScript Regular Expressions.

Για παράδειγμα, στην αναζήτηση της κανονικής έκφρασης για το πεδίο "startOfLinkRegEx" του wrapper επιπέδου 1, για την εξαγωγή των διευθύνσεων των άρθρων του antenna αντιμετωπίστηκε μία ανάλογη δυσκολία. Σε ορισμένες από τις διευθύνσεις προηγείται η έκφραση <a class="headlink" href=", σε άλλες η <a class="navlink" href="/ ή η ίδια πάλι με διαφορετική λέξη ανάμεσα στα «αυτάκια». Αντί, λοιπόν, να περαστούν όλες αυτές οι πιθανές εκφράσεις με κίνδυνο μάλιστα να παραλειφθούν ορισμένες ή να μην μπορούν να προβλεφθούν κάν, χρησιμοποιήθηκε η κανονική έκφραση <a class="*" href="/, στην οποία ο αστερίσκος (*) δηλώνεται ένας οποιοσδήποτε χαρακτήρας.

Ανάλογα πρόβλημα συναντήθηκε και στον ορισμό της κανονικής έκφρασης για το πεδίο "startOfLinkRegEx" του wrapper επιπέδου 1 της Le Monde, μόνο που αυτή τη φορά το πρόβλημα δημιουργούσαν οι αριθμοί. Σύμφωνα, λοιπόν, με τον κώδικα της Le Monde τα strings που προηγούνται των διευθύνσεων έχουν ως εξής: <a href=", <a href=" και κάθε άρθρο έχει ένα συνεχόμενο αριθμό. Ως λύση, λοιπόν, δόθηκε η κανονική έκφραση <a href=", όπου το "(\+|-)?[1-9][1-9]*(\.[1-9]*)?" ισούται με οποιοδήποτε αριθμός ανάμεσα στο 1 και στο 9.

Σε ορισμένες περιπτώσεις η χρήση πρότυπων κανονικών εκφράσεων δεν ήταν απαραίτητη, καθώς οι wrappers δούλεψαν το ίδιο αποδοτικά

ορίζοντάς τους απλά strings. Για παράδειγμα, στο πεδίο "startOfTitleRegEx" του wrapper 2 για το bbc.greek χρησιμοποιήθηκε το ίδιο αποτελεσματικά το string ώς έχει, δηλαδή: , όσο και με την αντικατάσταση του κενού από την κανονική έκφραση \s* (Σύμφωνα με το espresso): <div\s*class="storytext.

Στα πρώτα στάδια της υλοποίησης των wrappers η Βάση Περιεχομένου διατηρούσε ακόμα την αρχική της μορφή, όπως αυτή είχε σχεδιαστεί από τον φοιτητή Αλεξόπουλο Άγγελο. Στην πορεία, όμως, και κατά την εφαρμογή των κανονικών εκφράσεων στο σύστημα, παρουσιάστηκε ξεκάθαρη η ανάγκη αναδιοργάνωσης τόσο της Βάσης Περιεχομένου, όσο και σε πολλά σημεία του κώδικα του Συστήματος Εξατομικευμένης Ενημέρωσης. Η ανάγκη αυτή ήταν αποτέλεσμα των πηγών που προστέθηκαν στην παρούσα πτυχιακή εργασία, καθώς η αρχική μορφή της Βάσης είχε σχεδιαστεί έτσι ώστε να εφαρμόζεται στην δομή της Ναυτεμπορικής και του CNN.

Στις πρώτες δοκιμές του scanner, και ενώ η εξαγωγή ειδήσεων από την Ναυτεμπορική διεξαγόταν ομαλά, δημιουργήθηκαν προβλήματα με την εξαγωγή πληροφοριών από τις υπόλοιπες πηγές. Για το CNN οι wrappers δεν μπορούσαν να αντλήσουν τις διευθύνσεις των άρθρων, διότι ο κώδικας της πηγής αυτής είναι δομημένος κατά τέτοιο τρόπο ώστε να δυσκολεύει τον καθορισμό κανονικών εκφράσεων. Συγκεκριμένα, πρίν από την διεύθυνση κάθε άρθρου δεν παρέχεται ένα μοναδικό string το οποίο θα χρησιμοποιηθεί ως κανονική έκφραση για την εξαγωγή της ζητούμενης πληροφορίας. Στην **εικόνα 9** φαίνεται ένα κομμάτι του κώδικα της ιστοσελίδας http://www.cnn.com/POLITICS/. Το string το οποίο προηγείται των διευθύνσεων είναι ένα συνηθισμένο tag (<a href="/) το οποίο συναντάται πολλές φορές μέσα στον κώδικα. Για τον λόγο αυτό τελικά το CNN δεν μπόρεσε να συμπεριληφθεί στο Σύστημα Εξατομικευμένης Ενημέρωσης.

```
<div class="cnnBulletList"> &#8226; 
```

```
<a href="/2005/POLITICS/09/28/delay/index.html">Indicted DeLay defiant</a> |
<a href="/2005/POLITICS/09/28/blunt.dreier/index.html">House GOP fills leadership gap</a><br>&#8226; 
<a href="/2005/POLITICS/09/28/scotus.bush.ap/index.html">Bush narrows Supreme Court list</a> |
```

Εικόνα 9: Τμήμα του κώδικα της ιστοσελίδας: http://www.cnn.com/POLITICS/.

Εκτός από το CNN ανάλογα προβλήματα, αλλά όχι δυσεπίλυτα, αντιμετωπίστηκαν επίσης και με τις υπόλοιπες πηγές, οι οποίες εφαρμόστηκαν πρώτη φορά στο σύστημα. Ένα ζήτημα, το οποίο ανέκυψε, ήταν, ότι όλες οι πηγές αυτές (Ελευθεροτυπία, Antenna, bbc.greek, Le Monde) δεν έδιναν στον κώδικά τους απόλυτες διευθύνσεις. Όταν, λοιπόν, οι wrappers "έτρεχαν" τον κώδικα, οι διευθύνσεις που κατάφερναν να εξάγουν δεν ήταν οι σωστές. Ως λύση στο πρόβλημα αυτό, προστέθηκε στη Βάση Περιεχομένου στον πίνακα *html_wrapper* ένα καινούριο πεδίο το "relativePath" , το οποίο παρέχει στον wrapper το κομμάτι εκείνο της διεύθυνσης το οποίο λείπει. Για παράδειγμα στην Ελευθεροτυπία το "relativePath" είναι http://www.enet.gr/online/, σ' αυτό λοιπόν έρχεται να προσθέσει ο wrapper την πληροφορία-διεύθυνση που εξάγει και το αποτέλεσμα είναι η απόλυτη url.

Ένα ακόμη πρόβλημα ήταν ότι οι wrappers δεν μπορούσαν να εξάγουν την πρώτη πρόταση για το bbc.greek, τον Antenna και τη Le Monde. Αρχικά, οι wrappers για την εξαγωγή της πρώτης πρότασης χρησιμοποιούσαν τα πεδία **begregSentence** και **endregSentence** του πίνακα *content_wrapper*. Στο πεδίο **BegregSentence** δινόταν μία κανονική έκφραση, η οποία προηγούνταν της πρότασης, και στο **endregSentence** μία που ακολουθούσε. Το πρόβλημα εντοπίζονταν στο δεύτερο πεδίο καθώς συνήθως στον κώδικα την πρόταση ακολουθούσε ένα αρκετά συνηθισμένο tag. Για παράδειγμα, στον Antenna η πρόταση τελείωνε απλά με τα στοιχεία: ">. Μία πρώτη προσπάθεια επίλυσης του προβλήματος βασίστηκε στα Special JavaScript Regular Expressions.

Αν, λοιπόν, παρατηρήσει κανείς τον κώδικα του Antenna θα διαπιστώσει πως μετά το "> ακολουθεί ο τίτλος του άρθρου γραμμένος στα Ελληνικά:href="/articleDetail/0,3091,109291,00.html">Ανεστάλη ο νόμος για τον «βασικό μέτοχο». Προκειμένου, λοιπόν, να καταστήσουμε το περιεχόμενο του πεδίου μοναδικό, χρησιμοποιήθηκε η εξής κανονική έκφραση: ">\p{InGreek}, η οποία δηλώνει ότι η διεύθυνση εντοπίζεται πριν εκείνο ακριβώς το "> το οποίο ακολουθείται από Ελληνικό χαρακτήρα. Όμως, παρότι η κανονική έκφραση οδηγεί τον wrapper σωστά, το αποτέλεσμα ήταν πάντα το ίδιο: οι wrappers δεν κατέφερναν να εξάγουν την πρόταση. Η λύση, λοιπόν, που προτάθηκε και εφαρμόστηκε τελικά ήταν να πραγματοποιηθούν κάποιες αλλαγές στον κώδικα του συστήματος. Την αλλαγή στον κώδικα ανέλαβαν οι φοιτητές: Αλέξανδρος Μουζακίδης και Χρήστος Ντότσης.

Με τις αλλαγές που πραγματοποιήθηκαν στον κώδικα του συστήματος πολλά από τα πεδία της Βάση Περιεχομένου δεν χρειάζονταν πια. Έτσι, λοιπόν, καταργήθηκαν τα πεδία του *html_wrapper* "flagSentence", "notmatch", "flag1" και "flag2", τα οποία οδηγούσαν το wrapper 1 σε ένα χαρακτηριστικό σημείο στον κώδικα απ' όπου έπρεπε να ξεκινήσει την αναζήτηση των κανονικών εκφράσεων "startOfLinkRegEx" για την εξαγωγή των διευθύνσεων. Ουσιαστικά, λειτουργούσαν ως βοηθητικές κανονικές εκφράσεις. Από τον πίνακα content_wrapper καταργήθηκαν, επίσης, τα αντίστοιχα πεδία "flagSentence" και "flag", όπως και τα πεδία "BegregSentence" και "endregSentence", τα οποία έδειχναν στον wrappert 2 από πού να αρχίσει την εξαγωγή της πρώτης πρότασης του άρθρου. Πλέον, η διαδικασία αυτή πραγματοποιείται αυτόματα με εντολή του κώδικα του συστήματος.

Η κατάργηση των παραπάνω πεδίων έχει σαν αποτέλεσμα την ελλάτωση των κανονικών εκφράσεων, γεγονός που οδηγεί το σύστημα σε μεγαλύτερη ανεξαρτησία. Γιατί ο ασταθής χαρακτήρας των κανονικών εκφράσεων, πολύ συχνά δεν επιτρέπει την εξαγωγή ειδήσεων από έναν

ιστότοπο ο οποίος έχει αλλάξει την δομή του. Γεγονός άλλωστε το οποίο αποδείχτηκε με το CNN. Γενικότερα η χρήση των κανονικών εκφράσεων για την λειτουργία των wrappers δεν αποτελεί τον πλέον ασφαλή τρόπο εξαγωγής δεδομένων. Για το λόγο αυτό ήδη έχει ξεκινήσει μία προσπάθεια εκμετάλλευσης RSS, τα οποία παρέχουν μεγαλύτερη στεθερότητα στο σύστημα.

Στην επόμενη ενότητα παρουσιάζονται τα πεδία της Βάσης Περιεχομένου, όπως έχουν μετά τις αλλαγές του συστήματος από τους φοιτητές, Αλέξανδρο Μουζακίδη και Χρήστο Ντούτση.

5.6 Νέα Βάση Περιεχομένου (Content Database)

Η Βάση Περιεχομένου αποτελεί την κυρίως βάση δεδομένων του συστήματος, όπου φυλάσσονται οι ειδήσεις που εξάγονται από τον Content Scanner και οι κανονικές εκφράσεις που χρησιμοποιούνται από τα προγράμματα wrappers. Σε ένα πρώτο στάδιο η υλοποίηση των wrappers πραγματοποιήθηκε στη βάση περιεχομένου, η οποία είχε σχεδιαστεί από τον φοιτητή Αλεξόπουλο Άγγελο. Στη συνέχεια, όμως, για λόγους πρακτικότητας η Βάση Περιεχομένου άλλαξε ως προς την ονομασία των πεδίων και επίσης προστέθηκαν κάποια επιπλέον πεδία, ενώ καταργήθηκαν κάποια άλλα. Η διαδικασία αυτή πραγματοποιήθηκε από τους φοιτητές του Τμήματος Πληροφορικής του ΤΕΙ Αθηνών, Αλέξανδρο Μουζακίδη και Ντούτση Χρήστο.

Στην ενότητα αυτή παρουσιάζονται αναλυτικά καινούριοι οι πίνακες της Βάσης Περιεχομένου. Για κάθε πίνακα δίνονται τα ονόματα των πεδίων που τον αποτελούν, καθώς και μια σύντομη περιγραφή του ρόλου τους. Την Βάση αποτελούν έξι πίνακες οι οποίοι περιέχουν όλη την πληροφορία για τις κατηγορίες των ειδήσεων, τις πηγές των ειδήσεων και τις κανονικές εκφράσεις.

- **Πίνακας news_category**

Στον πίνακα αυτό φυλάσσονται οι κατηγορίες ειδήσεων που εξάγονται από τον Content Scanner.

news _category	Περιγραφή
News_id	Το κλειδί του πίνακα
News_Subject	Η κατηγορία ειδήσεων στα αγγλικά (π.χ. Politics)
News_Subject_ingreek	Η κατηγορία ειδήσεων στα ελληνικά (π.χ. Πολιτική)

- **Πίνακας source_name**

Στον πίνακα αυτό φυλάσσονται οι ειδησεογραφικές πηγές από τις οποίες εξάγονται ειδήσεις από τον Content Scanner.

Source_name	Περιγραφή
Source_id	Το κλειδί του πίνακα
Source_name	Το όνομα της πηγής (π.χ. Ναυτεμπορική)
Source_url	Η διεύθυνση στο Web (URL) της πηγής (π.χ. http://www.naftemporiki.gr)

- **Πίνακας source_category**

Στον πίνακα αυτό φυλάσσονται οι διευθύνσεις στο Web (URL) των ιστοσελίδων των πηγών για τις συγκεκριμένες κατηγορίες ειδήσεων από τις οποίες εξάγονται ειδήσεις από τον Content Scanner.

source_category	Περιγραφή
Id	Το κλειδί του πίνακα
url	Η διεύθυνση στο Web (URL) της πηγής για την συγκεκριμένη κατηγορία (π.χ. http://www.naftemporiki.gr/news/static/fin.htm)
CategoryId	Το κλειδί από τον πίνακα news_category, το οποίο αντιστοιχεί στην συγκεκριμένη κατηγορία
SourceId	Το κλειδί από τον πίνακα source_name, το οποίο

	αντιστοιχεί στην συγκεκριμένη πηγή
IsXml	Δηλώνεται αν κάποια Διεύθυνση είναι κωδικοποιημένη σε xml

- **Πίνακας news_content**

Στον πίνακα αυτό φυλάσσονται οι ειδήσεις όπως εξάγονται από τον Content Scanner

news_content	Περιγραφή
News_id	Το κλειδί του πίνακα
Title	Ο τίτλος της είδησης
Category	Το κλειδί της κατηγορίας του πίνακα news_category, στην οποία ανήκει η είδηση
Ndate	Η ημερομηνία εξαγωγής της είδησης
Source	Το κλειδί της πηγής του πίνακα source_name, στην οποία ανήκει η είδηση
url	Η διεύθυνση στο Web (URL) της είδησης
Sentence	Η πρώτη φράση της είδησης

- **Πίνακας html_wrapper**

Στον πίνακα αυτό φυλάσσονται οι πληροφορίες εισόδου του Content Scanner και συγκεκριμένα οι πληροφορίες που χρησιμοποιούν οι wrappers επιπέδου 1 για να εξάγουν τις διευθύνσεις των ειδήσεων που θα επεξεργασθούν οι wrappers επιπέδου 2.

Html_wrapper	Περιγραφή
Id	Το κλειδί του πίνακα
SourceId	Το κλειδί του πίνακα source_category
startOfLinkRegEx	Κανονική έκφραση που δείχνει από ποιο σημείο να αρχίσει η εξαγωγή πληροφορίας (δηλαδή του URL)
RelativePath	Το κομμάτι της διεύθυνσης που χρειάζεται για τον

	σχηματισμό της απόλυτης URL

- **Πίνακας content_wrapper**

Στον πίνακα αυτό φυλάσσονται οι πληροφορίες εισόδου του Content Scanner και συγκεκριμένα οι πληροφορίες που χρησιμοποιούν οι wrappers επιπέδου 2 για να εξάγουν τις πληροφορίες που αποθηκεύονται στον πίνακα news_content.

Content_wrapper	Περιγραφή
ContentGrabber	Το κλειδί του πίνακα
SourceId	Το κλειδί του πίνακα source_category
startOfTitleRegEx	Κανονική έκφραση που δείχνει από ποιο σημείο να αρχίσει η εξαγωγή του τίτλου της είδησης
EndOfTitleRegEx	Κανονική έκφραση που δείχνει σε ποιο σημείο να σταματήσει η εξαγωγή του τίτλου της είδησης
StartOfContentRegEx	Κανονική έκφραση που δείχνει από που να ξεκινήσει να εξάγεται το άρθρο της είδησης
EndOfContentRegEx	Κανονική έκφραση που δείχνει που να σταματήσει να εξάγεται το άρθρο της είδησης

Οι παραπάνω πίνακες αποτελούν τμήματα μίας βάσης Microsoft Access 2000.

Επίλογος

Κατά τη διάρκεια υλοποίησης της πτυχιακής εργασίας απαντήθηκαν διάφορα προβλήματα, τα οποία αποτέλεσαν έναυσμα για τη δημιουργία νέων πρακτικών πιο αποτελεσματικών, ιδιαίτερα σε ότι αφορά τους wrappers. Όπως επισημάνθηκε και προηγουμένως πολύ συχνά βρεθήκαμε αντιμέτωποι με το πρόβλημα της αλλαγής του κώδικα των ιστοσελίδων. Το φαινόμενο αυτό αποδείχτηκε αναπόφευκτο και απαιτεί την αλλαγή του τρόπου λειτουργίας του συστήματος σε ότι αφορά την εξαγωγή ειδήσεων.

Ήδη, λοιπόν, έχουν αρχίσει να πραγματοποιούνται κάποια βήματα προς αυτή την κατεύθυνση με την χρήση των RSS, η οποία φαίνεται να είναι πραγματικά αποτελεσματική.

Πέραν αυτού και σε σχέση με το Σύστημα Εξατομικευμένης Ενημέρωσης αλλαγές πραγματοποιήθηκαν και στην Διεπαφή Περιεχομένου (Content Presenter), προκειμένου να προστεθούν οι κατηγορίες των IPTC News Codes, σύμφωνα με τις οποίες μπορεί ο χρήστης να επιλέξει τις ειδήσεις που τον ενδιαφέρουν. Σύμφωνα με τις αλλαγές που έγιναν στο σύστημα με την προσθήκη νέων πηγών ειδησεογραφίας, η κατηγοριοποιήση των ειδήσεων γίνεται από τις κατηγορίες: Πολιτισμός, Οικονομία, Πολιτική, Επιστήμη και Τεχνολογία, Κοινωνικά ζητήματα, Αθλητισμός, Καιρός, Διεθνή, Ελλάδα.

Επιπλέον, αξίζει να σημειωθεί οτι παρά την ύπαρξη και την ανάπτυξη προτύπων οντολογιών και μεταδεδομένων για την ειδησεογραφία, οι περισσότερες από τις υπηρεσίες ηλεκτρονικής ενημέρωσης λειτουργούν σχετικά ανεξάρτητα σε οτι αφορά την κατηγοριοποίηση του υλικού τους. Μάλιστα, συνήθως η κατηγοριοποίηση πραγματοποιείται με βάση κάποιες γενικές θεματικές κατηγορίες, γι’ αυτό και πολύ συχνά εντοπίζουμε τις ίδιες κατηγορίες κάτω από διαφορετική ονομασία. Για παράδειγμα, άρθρα που σύμφωνα με τους IPTC News Codes ανήκουν στην κατηγορία *τέχνες, πολιτισμός και διασκέδαση*, στις ειδησεογραφικές πηγές οι κατηγορίες που τα περιελαμβάνουν τιτλοφορούνται άλλοτε ως *τέχνες,* άλλοτε ώς *πολιτισμός* ή ακόμα και ως *εκδηλώσεις, πολιτιστικά* και άλλα. Το φαινόμενο αυτό, ίσως να οφείλεται στο γεγονός οτι ο Ιστός ακόμα βρίσκεται σε ένα μεταβατικό στάδιο, όπου οι νέες τεχνολογίες ακόμα δεν βρίσκουν εφαρμογή σε όλες τις υπηρεσίες και σε όλες τις χώρες για οικονομικούς κυρίως λόγους.

Επίκεντρο των σημερινών εξελίξεων των νέων τεχνολογιών και ιδιαίτερα των υπηρεσιών ηλεκτρονικής ενημέρωσης αποτελεί ο χρήστης.
Στην παρούσα πτυχιακή εργασία μελετήθηκαν διάφορα πρότυπα, εφαρμογές και προγράμματα, τα οποία προωθούν την εξέλιξη των

υπηρεσιών ενημέρωσης, παρέχοντας στους αναγνώστες ευκολία στην αναζήτηση και ανάκτηση των πληροφοριών που τους ενδιαφέρουν. Άλλωστε, το όραμα του Σημασιολογικού Ιστού είναι να πραγματοποιείται η επεξεργασία δεδομένων με τέτοιο τρόπο, ώστε να ικανοποιούνται οι όλο και αυξανόμενες απαιτήσεις των χρηστών. Η χρήση, λοιπόν, προτύπων και νέων τεχνολογιών για την οργάνωση και διάθεση των πληροφοριών, που κυκλοφορούν στο Διαδίκτυο, αποτελεί επιτακτική ανάγκη σε μία κοινωνία όπου καθημερινά παράγεται τεράστιος όγκος νέων πληροφορίων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Antenna, http://news.antenna.gr/
2. Antoniou, G., F. van Harmelen. 2003, Web Ontology Language: OWL, Springer-Verlag
3. Antoniou, G., M. Baldoni, C. Baroglio, V. Patti, R. Baumgartner, T. Eiter, M. Herzog, R. Schindlauer, H. Tompits, F. Bry, S. Schaffert, N.

Henze, W. May. 2004, Reasoning Methods for Personalization on the Semantic Web. *Annals of Mathematics, Computing and Teleinformatics (AMCT) 1(2),* Institute for Informatics, Georg-August-Univesität Göttingen, pp. 1-24

4. Antoniou, G., E. Franconi, F. van Harmelen. Introduction to Semantic Web Ontology Languages

5. Baader, F., I. Horrocks and U. Sattler. 2005, Description Logics as Ontology Languages for the Semantic Web, Springer

6. BBC Greek, http://www.bbc.co.uk/greek/

7. Berners-Lee, T., J. Handler, and O. Lassila. 2001, The Semantic Web, *Scientific American*

8. Bonett, M. Personalization of Web Services: Oportunities and Challenges. *Ariadne,* available at http://ariadne.ac.uk/issue28/personalization/

9. CNN, http://www.cnn.com/

10. Decker, S., S. Malnik, F. van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, I. Horrocks. 2000, The Semantic Web: The Roles of XML and RDF. In *IEEE Internet Computing*

11. Fernandez-Garcia, N., L. Sanchez-Fernandez. 2004, Building an Ontology for NEWS Applications. *In proceedings of the 3rd International Semantic Web Conference ISWC2004,* http://iswc2004.semanticweb.org

12. Henze, N., Kriesell, M. 2004, Personalization Functionality for the Semantic Web: *Architectural Outline and First Sample Implementations,* Distributed Systems Institute – Knowledge Based Systems

13. Horrocks, I., P. F. Patel-Schneider, F. van Harmelen. 2002, Reviewing the Desing of DAML+OIL: *An Ontology Language for the Semantic Web,* American Association Intelligence, www.aaai.org

14. IEEE Internet Computing http://computer.org/internet

15. Information Interchange Model (*IIM*). Available at: http://www.iptc.org/IIM/

16. IPTC NewsML Web. Available at: http://www.newsml.org

17. Kalfoglou, Y., J. Domingue, E. Motta, M. Vargas-Vera, S. Buckingham Shum. MyPlanet: *An Ontology-driven Web-based personilised news service,* Knowledge Media Institute (KMi)
18. Kinecta: Industry Standards: PRISM http://www.kinecta.com/resources/prism.html
19. Kravatz, H. 2000, Desining Web Personalization Features
20. Liu, L., W. Han, D. Buttler, C. Pu, W. Tang.1999, An XML-based Wrapper Generator for Web Information Extraction, *in Proceedings of ACM SIGMOD International Conference on Management of Data*
21. Markellou, P., M. Rigou, S. Sirmakessis & A. Tsakalidis. 2004, Personalization in the Semantic Web Era: *a glance ahead*. *WIT Press*
22. Manola, F. 2002, The Semantic Web and the Role of Information Systems Research, *The MITRE Corporation.* Available at the Workshop Web site as: http://lsdis.cs.uga.edu/SemNSF/Manoloa-Position.doc
23. McIlraith, Sheila A., T.C. Son, H. Zeng. 2001, Semantic Web Services
24. Muslea, I. 1999, Extraction Patterns for Information Extraction Tasks: A Survey. In *IEEE Intelligent Systems,* available at http://www.ai.sri.com/~muslea/PS/ml4ie-aaai99.pdf
25. Neptuno Project (The). *Networked Semantic Team (NETS),* http://nets.ii.uam.es/neptuno
26. NEWS (*News Engine Web Services*) Home. Available at: http://www.news-project.com
27. NIFT: *News Industry Text Format.* Available at: http://www.nift.org
28. Noy, N. F., D. L. McGuiness. 2001, Ontology Development 101: *A guide to Creating Your First Ontology,* Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880
29. OWL, Web Ontology Language, WC3 Recommendation, Feb. 2004 http://www.w3.org/TR/owl-features/
30. Potock, T.E., M.T. Elmore, J.W. Reed, and N.F. Samatova. 2002, An Ontology-based HTML to XML Conversion Using Intelligent Agents, *In Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35'02)*
31. PRISM (Publishing Requirements for Industry Standard Metadata)

http://www.prismstandard.org/

32. RDF *Vocabulary Description Language1.0: RDF Schema.* Available at: http://www.w3.org/TR/rdf-schema

33. Sanfilippo, A., A. Bernardi, L. van Elst, L. S. Fernandez, M. Sintek. 2003, Position Paper: *Integrating Ontologies for Semantic Web Applications, ENABLER/ELSNET Workshop on International Roadmap for Languages Resources,* Paris

34. Shahabi, C., Yi-Shin Chen. Web Information Personalization: *Challenges and Approaches*

35. Shih-Hung, W., Tzing-Han Tsai and Wen-Lian Hsu. 2003, Domain Event Extraction and Representation with Domain Ontology, *Workshop on Information Intergration on the Web (IIWeb-03)*

36. Metadata: Subject Reference System and NewsML topicsets. Available at: http://www.iptc.org/metadata/

37. UMBC eBiquity
http://ebiquity.umbc.edu/us

38. Γεργατσούλης, Μ., Χ. Παπαθεοδώρου. Διαφάνειες μαθήματος Διαχείριση Γνώσης.

39. Ελευθεροτυπία, http://www.enet.gr/online/online
40. Ναυτεμπορική, http://www.naftemporiki.gr/
41. Τζουβάρας Β. Σημασιολογικός Ιστός, *5ο Σεμινάριο Κατάρτισης Multimine* (www.multimine.gr)

ΠΑΡΑΡΤΗΜΑ

Θεματικές κατηγορίες IPTC

ENGLISH	ΕΛΛΗΝΙΚΑ
Subjects / subject matters	**Θέματα** / θεματικές υποδιαιρέσεις-1ο επίπεδο

Arts, Culture & Entertainment	**01**	**Τέχνη, Πολιτισμός & Διασκέδαση**
Archeology	0001	Αρχαιολογία
Architecture	0002	Αρχιτεκτονική
Bull fighting	0003	Ταυρομαχίες
Carnival	0004	Καρναβάλι
Cinema	0005	Κινηματογράφος
Dance	0006	Χορός
Fashion	0007	Μόδα
Language	0008	Γλώσσα
Libraries & Museums	0009	Βιβλιοθήκες & Μουσεία
Literature	0010	Λογοτεχνία
Music	0011	Μουσική
Painting	0012	Ζωγραφική
Photography	0013	Φωτογραφία
Radio	0014	Ραδιόφωνο
Sculpture	0015	Γλυπτική
Television	0016	Τηλεόραση
Theatre	0017	Θέατρο
Crime, Law & Justice	**02**	**Έγκλημα, Νομοθεσία & Δικαιοσύνη**
Crime	0001	Έγκλημα
Judiciary	0002	Δικαστήριο
Police	0003	Αστυνομία
Punishment	0004	Κακοποίηση, Τιμωρία
Prison	0005	Φυλακή
Disasters & Accidents	**03**	**Καταστροφές & Ατυχήματα**
Drought	0001	Ξηρασία
Earthquake	0002	Σεισμός
Famine	0003	Λιμός
Fire	0004	Φωτιά
Flood	0005	Πλημμύρα
Industrial accident	0006	Βιομηχανικό ατύχημα
Meteorological disaster	0007	Καιρικές καταστροφές
Nuclear accident	0008	Πυρηνικές καταστροφές
Plague	0009	Επιδημία
Pollution	0010	Μόλυνση

Transport accident	0011	Ατυχήματα μεταφορών
Volcanic eruption	0012	Ηφαιστιακή έκρηξη
Economy, Business & Finance	**04**	**Οικονομία, Επιχειρήσεις & Δημοσιονομία**
Agriculture	0001	Γεωργία
Chemicals	0002	Χημικά προϊόντα
Computing & Information Technology	0003	Υπολογιστές & Τεχνολογία Πληροφόρησης
Construction & Property	0004	Κατασκευές & Ιδιοκτησία
Energy & Resources	0005	Ενέργεια & Πλουτοπαραγωγικές πηγές
Financial & Business Services	0006	Οικονομικές & Επιχειρηματικές υπηρεσίες
Goods Distribution	0007	Εμπορική διανομή
Macro Economics	0008	Μακροοινομία
Markets	0009	Αγορά
Media	0010	Μέσα Ενημέρωσης
Metal Goods & Engineering	0011	Εμπόριο μετάλλων & Μηχανολογία
Metals & Minerals	0012	Μέταλλα & Ορυκτά
Process Industries	0013	Μεταποιητικές βιομηχανίες
Tourism & Leisure	0014	Τουρισμός & Ελεύθερος χρόνος
Transport	0015	Μεταφορές
Education	**05**	**Εκπαίδευση**
Adult Education	0001	Εκπαίδευση ενηλίκων
Further Education	0002	Συνεχιζόμενη εκπαίδευση
Parent Organizations	0003	Εκπαιδευτικοί Οργανισμοί
Preschooling	0004	Προσχολική εκπαίδευση
Schools	0005	Σχολεία, Κολλέγια & Ινστιτούτα
Teachers Unions	0006	Ενώσεις εκπαιδευτικών
University	0007	Πανεπιστήμιο
Environmental Issues	**06**	**Περιβαλλοντικά θέματα**
Alternative Energy	0001	Εναλλακτικές πηγές ενέργειας
Conservation	0002	Προστασία περιβάλλοντος
Energy Savings	0003	Εξοικονόμηση ενέργειας
Environmental Politics	0004	Περιβαλλοντική πολιτική
Environmental pollution	0005	Μόλυνση του περιβάλλοντος
Natural resources	0006	Φυσικοί πόροι

Nature	0007	Φύση
Population	0008	Πληθυσμός
Waste	0009	Απόβλητα
Water Supplies	0010	Προμήθεια νερού
Health	**07**	**Υγεία**
Diseases	0001	Νοσήματα
Health treatment	0002	Θεραπευτική αγωγή
Health organizations	0003	Οργανισμοί υγείας
Medical research	0004	Ιατρική έρευνα
Medical staff	0005	Ιατρικό προσωπικό
Medicines	0006	Φάρμακα
Preventative medicine	0007	Προληπτική ιατρική
Human Interest	**08**	**Ανθρωπιστικά ενδιαφέροντα**
Animals	0001	Ζώα
Curiosities	0002	Αξιοπερίεργα
People	0003	Άνθρωποι
Labour	**09**	**Εργασία**
Apprentices	0001	Επαγγελματική εκπαίδευση
Collective contracts	0002	Συλλογικές συμβάσεις
Employment	0003	Απασχόληση
Labour dispute	0004	Εργατικές διεκδικήσεις
Labour legislation	0005	Εργατική νομοθεσία
Retirement	0006	Συνταξιοδότηση
Retraining	0007	Επανεκπαίδευση
Strike	0008	Απεργία
Unemployment	0009	Ανεργία
Unions	0010	Συνδικάτα
Wages & Pensions	0011	Αμοιβές και επιδόματα
Work Relations	0012	Εργασιακές σχέσεις
Lifestyle & Leisure	**10**	**Τρόπος ζωής & Ελεύθερος χρόνος**
Games	0001	Παιχνίδια
Gaming & Lotteries	0002	Τυχερά παιχνίδια & Λαχεία
Gastronomy	0003	Γαστρονομία

Hobbies	0004	Χόμπυ
Holidays or vacations	0005	Διακοπές
Tourism	0006	Τουρισμός
Politics	**11**	**Πολιτική**
Defence	0001	Άμυνα
Diplomacy	0002	Διπλωματία
Elections	0003	Εκλογές
Espionage & Intelligence	0004	Κατασκοπία
Foreign Aid	0005	Ξένη βοήθεια
Government	0006	Κυβέρνηση
Human Rights	0007	Ανθρώπινα δικαιώματα
Local authorities	0008	Τοπικές αρχές
Parliament	0009	Κοινοβούλιο
Parties	0010	Κόμματα
Refugees	0011	Πρόσφυγες
Regional authorities	0012	Περιφερειακές αρχές
State Budget	0013	Κρατικός προϋπολογισμός
Treaties & Organisations	0014	Διαπραγματεύσεις και Οργανισμοί
Religion & Belief	**12**	**Θρησκεία και Πίστη**
Cults & sects	0001	Λατρείες και Αιρέσεις
Faith	0002	Πίστη
Free masonry	0003	Μασονία
Religious institutions	0004	Θρησκευτικοί θεσμοί
Science & Technology	**13**	**Επιστήμη και Τεχνολογία**
Applied Sciences	0001	Εφαρμοσμένες επιστήμες
Engineering	0002	Μηχανολογία
Human Sciences	0003	Ανθρωπιστικές επιστήμες
Natural Sciences	0004	Φυσικές επιστήμες
Philosophical Sciences	0005	Φιλοσοφία
Research	0006	Έρευνα
Scientific exploration	0007	Επιστημονικές ανακαλύψεις
Space programmes	0008	Διαστημικά προγράμματα
Social Issues	**14**	**Κοινωνικά ζητήματα**

Addiction	0001	Εξάρτηση
Charity	0002	Φιλανθρωπία
Demographics	0003	Δημογραφία
Disabled	0004	Άτομα με ειδικές ανάγκες
Euthanasia	0005	Ευθανασία
Family	0006	Οικογένεια
Family planning	0007	Οικογενειακός προγραμματισμός
Health insurance	0008	Ασφάλειες ζωής
Homelessness	0009	Άστεγοι
Minority groups	0010	Μειονότητες
Pornography	0011	Πορνογραφία
Poverty	0012	Φτώχεια
Prostitution	0013	Πορνεία
Racism	0014	Ρατσισμός
Welfare	0015	Κοινωνική πρόνοια

Sport	**15**	**Αθλητισμός**
Aero and Aviation Sports	0001	Αεροβατικά και Αεροπορικά Αθλήματα
Alpine Skiing	0002	Αλπινικό σκι
American Football	0003	
Archery	0004	Τοξοβολία
Athletics, Track & Field	0005	Κλασικός αθλητισμός
Badminton	0006	
Baseball	0007	
Basketball	0008	Μπάσκετ
Biathlon	0009	Δίαθλο (σκι και σκοποβολή)
Billiards, Snooker and Pool	0010	
Bobsleigh	0011	
Bowling	0012	Μπόουλινγκ
Bowls & Petanque	0013	
Boxing	0014	Πυγμαχία
Canoeing & Kayaking	0015	Κανό & Καγιάκ
Climbing	0016	Ορειβασία
Cricket	0017	
Curling	0018	
Cycling	0019	Ποδηλασία
Dancing	0020	Αθλητικός Χορός

Diving	0021	Καταδύσεις
Equestrian	0022	Ιππικοί αγώνες
Fencing	0023	Ξιφασκία
Field Hockey	0024	
Figure Skating	0025	Καλλιτεχνικό πατινάζ
Freestyle Skiing	0026	Σκι
Golf	0027	Γκόλφ
Gymnastics	0028	Γυμναστική
Handball (Team)	0029	Χάντμπωλ
Horse Racing, Harness Racing	0030	Ιππασία
Ice Hockey	0031	
Jai Alai (Pelota)	0032	
Judo	0033	Τζούντο
Karate	0034	Καράτε
Lacrosse	0035	
Luge	0036	
Marathon	0037	Μαραθώνιος
Modern Pentathlon	0038	Μοντέρνο πένταθλο
Motor Racing	0039	Μηχανοκίνητοι αγώνες
Motor Rallying	0040	Αγώνες αυτοκινήτου
Motorcycling	0041	Αγώνες μηχανών
Netball	0042	
Nordic Skiing	0043	
Orienteering	0044	
Polo	0045	
Power Boating	0046	
Rowing	0047	
Rugby League	0048	
Rugby Union	0049	
Sailing	0050	Ιστιοπλοΐα
Shooting	0051	Κυνήγι
Ski Jumping	0052	
Snow Boarding	0053	
Soccer	0054	Ποδόσφαιρο
Softball	0055	
Speed Skating	0056	
Speedway	0057	

Sports Organisations	0058	Αθλητικές ενώσεις
Squash	0059	
Sumo Wrestling	0060	
Surfing	0061	Σέρφινγκ
Swimming	0062	Κολύμβηση
Table Tennis	0063	Πίνγκ-Πόνγκ
Taekwon-Do	0064	Τέικβον-ντο
Tennis	0065	Τένις
Triathlon	0066	Τρίαθλον
Volleyball	0067	Βόλεϋ
Water Polo	0068	Πόλο
Water Skiing	0069	Θαλάσσιο σκί
Weight lifting	0070	Άρση βαρών
Windsurfing	0071	Ιστιοσανίδα
Wrestling	0072	Πάλη

Unrest, Conflicts & War	**16**	**Αναταραχές, Διαμάχες και Πόλεμος**
Acts of terror	0001	Τρομοκρατία
Armed conflict	0002	Ένοπλες συγκρούσεις
Civil unrest	0003	Πολιτισμικές αναταραχές
Coup d'Etat	0004	Πραξικόπημα
Guerrilla activities	0005	Δραστηριότητες ανταρτών
Massacre	0006	Σφαγές
Riots	0007	Ταραχές
Violent demonstrations	0008	Βίαιες διαδηλώσεις
War	0009	Πόλεμος

Weather	**17**	**Καιρός**
Forecasts	0001	Προβλέψεις
Global change	0002	Παγκόσμιες αλλαγές
Reports	0003	Αναφορές
Statistics	0004	Στατιστικές
Warnings	0005	Προειδοποιήσεις