Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακή Διατριβή
στα Πληροφοριακά και Επικοινωνιακά Συστήματα

Παρακολούθηση της Εξέλιξης των Αντικειμένων

Ευαγγελία Σ. Ρουτζούνη

Επιβλέπουσα Καθηγήτρια
Αικατερίνη Ιωάννου

Μάιος 2013

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Παρακολούθηση της Εξέλιξης των Αντικειμένων

Ευαγγελία Σ. Ρουτζούνη

Επιβλέπουσα Καθηγήτρια

Αικατερίνη Ιωάννου

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε
προς μερική εκπλήρωση των απαιτήσεων για απόκτηση

μεταπτυχιακού τίτλου σπουδών
στα Πληροφοριακά και Επικοινωνιακά Συστήματα

από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών
του Ανοικτού Πανεπιστημίου Κύπρου

Μάιος 2013

Περίληψη

Κατά τη διάρκεια των τελευταίων ετών η ανάρτηση και τροποποίηση δεδομένων στον παγκόσμιο ιστό από μη έμπειρους χρήστες έχει καταστήσει τις περιγραφές των αντικειμένων ευμετάβλητες. Η παρούσα μεταπτυχιακή διατριβή μελέτησε και ανέπτυξε ένα σύστημα (εφαρμογή ιστού WOE) που, με απλό και κατανοητό τρόπο, επιτρέπει στους χρήστες του (ανθρώπους ή και άλλα συστήματα) να αναζητούν/μελετούν τις αλλαγές που έχουν γίνει σε ομάδες αντικειμένων ή στα χαρακτηριστικά ενός συγκεκριμένου αντικειμένου κατά την διάρκεια του χρόνου, πράγμα που συνιστά καινοτομία στο πεδίο. Τα συμπεράσματα της μελέτης των αλλαγών αυτών μπορούν να αξιοποιηθούν για τη βελτίωση αλγόριθμων επεξεργασίας ευμετάβλητων αντικειμένων με απώτερο στόχο τη βελτίωση της εμπειρίας του χρήστη στο Διαδίκτυο.

Με τη χρήση του συστήματος αυτού μελετήθηκαν τρεις εκδόσεις της Αγγλικής Wikipedia (3.5.1-dumped 3/2010, 3.6-dumped 11/2011, 3.8-dumped 6/2012), η οποία συγκαταλέγεται στις εφαρμογές ιστού όπου χρήστες έχουν ενεργό ρόλο στην παραγωγή των δεδομένων και άρα προσφέρεται για μελέτη ευμετάβλητων αντικειμένων.

Μερικές βασικές διαπιστώσεις από την εφαρμογή της μελέτης σε δεδομένα που περιγράφουν πρόσωπα στην Wikipedia είναι οι ακόλουθες:

- Εκθετικές αλλαγές στους ρυθμούς αύξησης των προσώπων και των ιδιοτήτων τους με ισχυρότερο το ρυθμό αύξησης για τα πρόσωπα - κατόχους κυβερνητικών θέσεων
- Ισχυρές διαφοροποιήσεις στο πλήθος μελών συγκεκριμένων ομάδων προσώπων - κατηγοριών (π.χ. οι μουσικοί, κυβερνητικοί, στρατιωτικοί είναι οι πολυπληθέστερες κατηγορίες σε όλες τις Wikipedia εκδόσεις)
- Σημαντική είσοδος νέων ιδιοτήτων στο σύστημα αλλά και ανάγκη προτυποποίησης του τρόπου απόδοσής τους για αποφυγή των πολλαπλών αναφορών
- Αλλαγές στα χαρακτηριστικά των προσώπων, που αρκετές φορές δεν είναι σημαντικές αλλά αφορούν μικρο-μεταβολές στην ακολουθία των χαρακτήρων που αποδίδει την τιμή (π.χ. ένα κενό περισσότερο, ένα underscore λιγότερο)
- Καταχωρήσεις που, ενώ δεν είναι πρόσωπα στον πραγματικό κόσμο, εν τούτοις είναι μοντελοποιημένα ως πρόσωπα στην Wikipedia (π.χ. "Presidency of Bill Clinton")
- Πρόσωπα σταθερά πιο δημοφιλή, με την έννοια ότι συγκεντρώνουν τις περισσότερες ιδιότητες από άλλα στην κατηγορία τους (π.χ. Arnold_ Schwarzenegger "ο δημοφιλέστερος ηθοποιός" , George H.W.. Bush "ο δημοφιλέστερος" πρόεδρος των Η.Π.Α)

Summary

During the last years post and modify data on the web from non-experienced users has made the descriptions of objects volatile.

This Master Thesis has implemented a system (web application WOE) which, in a simple and understandable way, enables users to look for/to study the changes made over time in the characteristics of a particular object (person) or in the characteristics of a group of objects. This consists an innovation in the field. The conclusions of such a study could be used for the improvement of algorithms processing volatile data, improving this way the total experience of the Web user.

One of the web applications where users have active role in the production of data - and therefore lends itself to the study of volatile items - is Wikipedia. Using the WOE system, a pilotic study on three versions of the English Wikipedia (3.5.1-dumped 3/2010, 3.6-dumped 11/2011, 3.8-dumped 6/2012) has been conducted. Some of the conclusions follow:

• Exponential changes in the growth rate of persons and their properties from 3/2010 to 6/2012 (the strongest growth rate for Wikipedia category "office-holders")

• Strong variations in the number of members of certain categories of people (categories "musical artists", "office-holders", "military_persons" are heavily populated in all Wikipedia versions)

• Significant number of properties' new entrances over time and a strong need for standardization in this field in order for multiple references to be avoided.

• Changes in persons' characteristics that aren't always meaningful, as they are changes of minor importance in the string sequence of a property's value. (e.g a space more, an underscore less)

• Several Wikipedia entries that are modelling as "persons", things that are not persons in the real world. (e.g. "Presidency of Bill Clinton", "82th Delaware General Assembly")

• Persons that are consistently more popular, in the sense that they have a greater number of properties than others in their category (e.g. Arnold_ Schwarzenegger most "popular" actor, George H.W. Bush most "popular president")

Ευχαριστίες

Θα ήθελα να ευχαριστήσω:

- τον Ακαδημαϊκό Υπεύθυνο του Μεταπτυχιακού Προγράμματος Σπουδών "Πληροφοριακά και Επικοινωνιακά Συστήματα" του Ανοιχτού Πανεπιστημίου Κύπρου, καθηγητή κ. Θανάση Χατζηλάκο, για την πολύτιμη συμπαράστασή του σε ό,τι κι αν χρειάστηκα κατά την διάρκεια της φοίτησής μου,

- την επιβλέπουσα καθηγήτρια κ. Αικατερίνη Ιωάννου για το ενδιαφέρον θέμα, την συνέπεια και την οργανωτικότητά της και

- τον καθηγητή κ. Βασίλειο Βερύκιο για την γνωριμία μου, τον Οκτώβρη του 2012, με τη γλώσσα Προγραμματισμού Python.

Θα ήθελα τέλος να εκφράσω την ευγνωμοσύνη μου :
στους γονείς μου-χωρίς αυτούς δεν θα ήμουν εδώ-
στον άνδρα μου-χωρίς την αγάπη του δεν θα είχα την δύναμη-
και
να αφιερώσω τη μεταπτυχιακή αυτή διατριβή στους δύο λατρεμένους γιους μου που είναι και θα παραμείνουν το μεγαλύτερο επίτευγμα της ζωής μου.

Λίλα Ρουτζούνη
Μάιος 2013.

Περιεχόμενα

Κεφάλαιο 1
Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιαστεί το πρόβλημα που θα μελετηθεί (1.1), η προτεινόμενη λύση (1.2) και η συνεισφορά της παρούσας μεταπτυχιακής διατριβής (1.3), ενώ στο 1.4 παρουσιάζεται η οργάνωση του παρόντος κειμένου.

1.1 Περιγραφή του προβλήματος.

Κατά τη διάρκεια των τελευταίων ετών, ο παγκόσμιος Ιστός φιλοξενεί δεδομένα μιας νέας μορφής που προέρχονται από μη έμπειρους χρήστες, οι οποίοι μπορούν να τα αναρτούν και να τα τροποποιούν. Έτσι τα δεδομένα αυτά είναι ήμι-δομημένα και εξαιρετικά ανομοιογενή. Τέτοια δεδομένα έχουν διαφορετικές απαιτήσεις διαχείρισης, οι οποίες δεν ικανοποιούνται αποδοτικά από υφιστάμενες υπηρεσίες.

Ως παράδειγμα ας λάβουμε ένα σύστημα διαχείρισης βάσεων δεδομένων, ΣΔΒΔ, (database management system). Ένα ΣΔΒΔ αντιστοιχεί σε ένα λογισμικό που επιτρέπει τη δημιουργία, διαχείριση και συντήρηση μιας συλλογής από σχετιζόμενα δεδομένα. Τα ΣΔΒΔ θεωρούν όμως πως όλα τα δεδομένα έχουν την ίδια σημασία/αξία. Γι' αυτό και ένα ευρετήριο θα χρειαστεί τον ίδιο χρόνο για αναζήτηση τόσο ενός δεδομένου Α όσο και ενός δεδομένου Β. Αν όμως είναι

γνωστό πως το δεδομένο Α είναι περισσότερο δημοφιλές από το Β, μπορεί να τροποποιηθεί το ευρετήριο, έτσι ώστε να επιστρέφει γρηγορότερα το Α από το Β.

Ως δεύτερο παράδειγμα ας λάβουμε ένα σύστημα που έχει σκοπό την ενοποίηση δεδομένων από διαφορετικές εφαρμογές. Τέτοια συστήματα ενσωματώνουν διάφορους αλγορίθμους (για βελτίωση της ενοποίησης δεδομένων και για την αύξηση της απόδοσης στην αναζήτηση) που στοχεύουν στον καθαρισμό (data cleaning) και στη σύζευξη των δεδομένων που περιγράφουν την ίδια οντότητα στον πραγματικό κόσμο. Όταν ο όγκος των δεδομένων είναι μεγάλος, το σύστημα καλείται να αποφασίσει ποια δεδομένα πρέπει να στείλει σε ποιους αλγορίθμους. Αν, λοιπόν, μπορεί να γνωρίζει την συχνότητα αλλαγής των δεδομένων, μπορεί καλύτερα να λάβει αυτή την απόφαση. Μπορεί για παράδειγμα να αποφασίζει ότι τα δεδομένα που αλλάζουν συχνά πρέπει να γίνονται αντικείμενο επεξεργασίας από τους αλγορίθμους του συστήματος και ότι τα δεδομένα που αλλάζουν σπάνια δεν πρέπει να αποστέλλονται για επεξεργασία.

1.2 Προτεινόμενη Λύση

Αυτή η μεταπτυχιακή διατριβή, έχοντας ως παράδειγμα τη Wikipedia, που είναι μια εφαρμογή ιστού στην οποία οι χρήστες έχουν ενεργό ρόλο στην παραγωγή των δεδομένων, έχει σκοπό αφ' ενός τη δημιουργία ενός εργαλείου μελέτης της εξέλιξης αντικειμένων ή/και ομάδων αντικειμένων που περιγράφουν οντότητες στον πραγματικό κόσμο και αφ'ετέρου την εξαγωγή συμπερασμάτων σχετικά με την εξέλιξή τους με χρήση του εργαλείου αυτού.

Συγκεκριμένα, δεδομένα από διαφορετικές εκδόσεις της Αγγλικής Wikipedia, τα οποία εμφανίζονται στο σύστημα της σε διάφορες χρονικές περιόδους (που αντιστοιχούν στα μεσοδιαστήματα των διάφορων αναθεωρήσεων των εκδόσεών της), θα αποθηκευτούν τοπικά και θα τύχουν επεξεργασίας προκειμένου να απομονωθούν εκείνα που αφορούν πρόσωπα. Τα τελευταία αυτά θα υποστούν νέα επεξεργασία με στόχο τη δημιουργία μιας «χρονικά ευαίσθητης» Βάσης Δεδομένων Οντοτήτων, η οποία θα περιέχει οντότητες-πρόσωπα με τις ιδιότητες, τις σχέσεις και τα χαρακτηριστικά τους στις διαφορετικές αυτές χρονικές περιόδους και, κατά συνέπεια, θα παρέχει τη δυνατότητα σύγκρισης ανάμεσα στα διαφορετικά «στιγμιότυπα» κάθε οντότητας στις περιόδους αυτές.

Τα παραπάνω θα υλοποιηθούν μέσω ενός συστήματος με μορφή εφαρμογής ιστού (WebApp), που θα επιτρέπει στους ενδιαφερομένους (χρήστες ή άλλα συστήματα) να εισάγουν, να

επεξεργάζονται και να μελετούν δεδομένα από Wikipedia εκδόσεις της επιλογής τους (custom use). Εναλλακτικά το σύστημα αυτό θα επιτρέπει στους ενδιαφερομένους να μελετήσουν τα αποτελέσματα της επεξεργασίας των δεδομένων από τις συγκεκριμένες Wikipedia εκδόσεις, που επιλέχθηκαν να τύχουν επεξεργασίας στα πλαίσια αυτής της μεταπτυχιακής διατριβής (default use).

Πρέπει να τονιστεί ότι δε θα διεξαχθεί στατιστική επεξεργασία των αποτελεσμάτων (όπου αυτά είναι επιδεκτικά σε τέτοιου είδους επεξεργασία), γιατί κεντρικός στόχος μας δεν είναι τόσο η ενδελεχής μελέτη των αποτελεσμάτων για εξαγωγή συμπερασμάτων (καθώς αυτά μπορεί να ποικίλουν ανάλογα με το χρήστη, σύστημα ή πρόσωπο που χρησιμοποιεί κάθε φορά το εργαλείο) όσο η πρόταση ενός μελετητικού μοντέλου.

1.3 Συνεισφορά της μεταπτυχιακής διατριβής

Η συνεισφορά της διατριβής είναι η δημιουργία ενός εργαλείου μελέτης της εξέλιξης αντικειμένων ή/και ομάδων αντικειμένων που περιγράφουν οντότητες στον πραγματικό κόσμο και με χρήση αυτού, η εξαγωγή συμπερασμάτων σχετικά με την εξέλιξή των αντικειμένων. Πιο συγκεκριμένα, τα κύρια σημεία της συνεισφοράς είναι:

- Η δημιουργία συστήματος, με μορφή εφαρμογής ιστού, με τη χρήση του οποίου θα δοθεί δυνατότητα σύγκρισης «των στιγμιότυπων» κάθε οντότητας, αλλά και των διαφορετικών ΒΔ σε διαφορετικές χρονικές περιόδους, κάτι που αποτελεί καινοτομία στο πεδίο.
- Ενδείξεις για το πώς χαρακτηριστικά γνωρίσματα προσώπων και ομάδων προσώπων μεταβάλλονται και συμπεράσματα π.χ. για τα δημοφιλέστερα πρόσωπα, τις δημοφιλέστερες ομάδες προσώπων ή για το ποιες κατηγορίες προσώπων είναι πολυπληθέστερες.
- Η προτεινόμενη χρήση των αποτελεσμάτων της μεταπτυχιακής διατριβής από αλγορίθμους για τη βελτίωση της λειτουργίας τους σε τέτοια ευμετάβλητα δεδομένα, όπως έχουμε περιγράψει και στο 1.1, αλλά και από άλλους χρήστες (συστήματα ή πρόσωπα) για περαιτέρω μελετητική δουλειά.

1.4 Οργάνωση κειμένου

Στο κεφάλαιο 1 παρουσιάζονται εισαγωγικές έννοιες που περιγράφουν το πρόβλημα, την προτεινόμενη λύση και τη συνεισφορά σε αυτό της παρούσας μεταπτυχιακής διατριβής.
Στο κεφάλαιο 2 παρουσιάζονται και σχολιάζονται εν συντομία σχετικές εργασίες.
Στο κεφάλαιο 3 δίνονται ορισμοί των εννοιών και περιγράφεται το μοντέλο της εξέλιξης των αντικειμένων.
Στο κεφάλαιο 4 παρουσιάζεται η μεθοδολογία που ακολουθήθηκε για την εξόρυξη των προσώπων και την επεξεργασία των δεδομένων τους.
Στο κεφάλαιο 5 παρουσιάζεται η αρχιτεκτονική και οι λειτουργίες του συστήματος το οποίο δημιουργήθηκε στα πλαίσια της παρούσας μεταπτυχιακής διατριβής, με σκοπό την παροχή δυνατότητας παρακολούθησης της εξελικτικής πορείας των αντικειμένων, καθώς και τα συμπεράσματα που προέκυψαν από μια πρώτη μελέτη των δεδομένων που επεξεργαστήκαμε.
Στο κεφάλαιο 6 συνοψίζονται τα συμπεράσματα και τα πεδία εφαρμογής και αξιοποίησής τους στην πράξη.
Στο κεφάλαιο 7 παρουσιάζεται η βιβλιογραφία που χρησιμοποιήθηκε.

Στα παραρτήματα:
στο Α παρατίθενται οδηγίες εγκατάστασης
στο Β:

στο Β1 τα SQLs που δημιουργίας της Temporal Entities' Data Base

στο Β2 τα SQLs για την εξόρυξη των τιμών των ιδιοτήτων στη TEDB

στο Γ η διαδικασία εισαγωγής επιπλέον ερευνητικών ερωτημάτων.

Κεφάλαιο 2

Σχετικές Εργασίες

Στις μέρες μας ο τυπικός χρήστης του Διαδικτύου δεν είναι πλέον θεατής που καταναλώνει πληροφορίες οι οποίες διατίθενται από κάποιο πάροχο. Μέσα από κοινωνικές εφαρμογές Web, όπως το MySpace, Blogosphere, Facebook, και μέσω υπηρεσιών, όπως το Twitter, ο χρήστης αποκτά έναν ενεργό ρόλο στην παραγωγή δεδομένων [3]. Μια εφαρμογή παρόμοιας λογικής είναι και η Wikipedia, μια διαδικτυακή εγκυκλοπαίδεια της οποίας τα λήμματα δημιουργούνται και συντηρούνται χάρη στη συνεργασία και προσφορά των χρηστών της, οι οποίοι έχουν στη διάθεσή τους τεχνολογίες που τους επιτρέπουν, όπως και στους χρήστες των κοινωνικών δικτύων, να παράγουν και να επεξεργάζονται δεδομένα χωρίς ιδιαίτερη κατάρτιση στις ΤΠΕ.

Η χρήση συστημάτων που περιγράφουν ή περιλαμβάνουν αντικείμενα δημιουργεί μια νέα πραγματικότητα στο διαδίκτυο και αυτή τη στιγμή υπάρχει στην επιστημονική κοινότητα μια ανοιχτή συζήτηση για τις επιπτώσεις της, αλλά και τους νέους τρόπους χειρισμού της.

Η υπάρχουσα αντιμετώπιση των ιστοσελίδων ως "δοχείων με λέξεις" [1] μπορεί να ήταν ικανοποιητική, όταν το διαδίκτυο ήταν ακόμα στην αρχή του, αλλά δεν είναι, κατά την άποψή μας ικανοποιητική πια, μια και του στερεί αυτό για το οποίο ακριβώς είναι πολύτιμο: τη δυνατότητα παροχής πληροφορίας σχετικής με το μεγάλο εύρος εννοιών (concepts) που περιλαμβάνει.

Συμφωνούμε με άλλους ερευνητές [1,3,4] ότι η εξαγωγή και ανάλυση των δεδομένων με τρόπο που να εστιάζει στις έννοιες αποτελεί το επόμενο στάδιο μετασχηματισμού του διαδικτύου, που θα αλλάξει τον τρόπο με τον οποίο αντλούμε πληροφορίες από αυτό, με τεράστιες κοινωνικές και οικονομικές συνέπειες. Δε θα βελτιωθεί μόνο ο τρόπος με τον οποίο οι μηχανές αναζήτησης αναγνωρίζουν και κατατάσσουν σχετικό περιεχόμενο, αλλά επίσης "θα δοθεί η δυνατότητα να υποστηριχθούν κριτήρια αναζήτησης περισσότερο πλούσια, προσανατολισμένα σε χαρακτηριστικά" [1], και να παραχθούν αποτελέσματα που να είναι συνθέσεις πληροφορίας βασισμένης σε έννοιες, η οποία δεν θα είχε παραχθεί με την συγκεκριμένη μορφή, αν οι έννοιες αυτές παρέμεναν διασκορπισμένες σε πολλές και διαφορετικές ιστοσελίδες.

Οι σημαντικές συνέπειες που μπορεί να έχει στη βελτίωση της εμπειρίας του χρήστη ένα διαδίκτυο των εννοιών (και επομένως και οντοτήτων) έχουν περιγραφεί αναλυτικά[1]. Ενδεικτικά αναφέρουμε session optimization, βελτιστοποίηση περιήγησης, αναζήτηση προσαρμοσμένη στην έννοια. Έχει επίσης καταδειχθεί η σημασία της διασύνδεσης των οντοτήτων με παραδείγματα εφαρμογής σε συγκεκριμένα επιστημονικά πεδία (ιστορικές σπουδές, βιολογίας) [4], ενώ υπάρχουν περιγράφες πρακτικών εφαρμογών του διαδικτύου εννοιών όπως π.χ. εξόρυξη πληροφορίας από κείμενο [5].

Συνήθεις τύποι εννοιών είναι οι οντότητες (entities), τα συμβάντα (events) και τα θέματα (topics) [1], αλλά εμείς στην παρούσα μεταπτυχιακή διατριβή θα εστιάσουμε στις οντότητες.

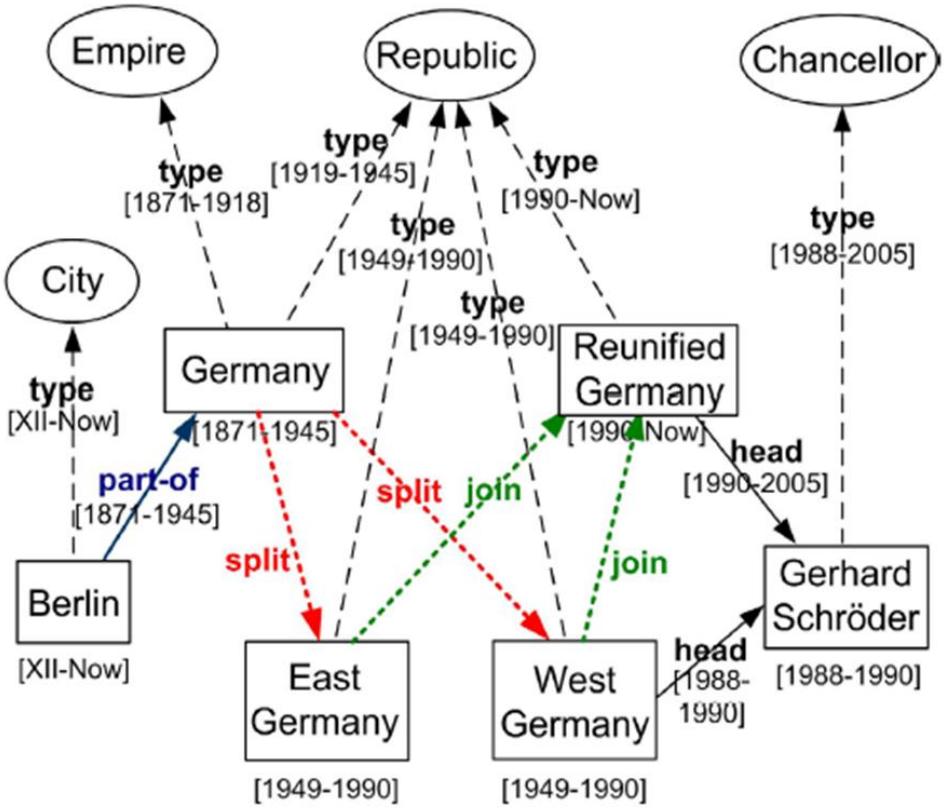
Η οντότητα είναι ένα σχεδιαστικό κατασκεύασμα που χρησιμοποιείται για να παραστήσει ένα αντικείμενο του πραγματικού κόσμου. Παράδειγμα μιας "προηγμένης" διαδικτυακής εφαρμογής, ήδη χτισμένης γύρω από οντότητες, είναι η Wikipedia.

Όπως όμως τα αντικείμενα του πραγματικού κόσμου, τα οποία οι οντότητες αναπαριστούν, αλλάζουν, έτσι και οι οντότητες δεν παραμένουν στατικές στο χρόνο: εξελίσσονται, συγχωνεύονται, διασπώνται, δημιουργούνται και εξαφανίζονται και "ενώ σημαντική ερευνητική προσπάθεια στο χώρο των βάσεων δεδομένων έχει κατευθυνθεί προς χρονικά μοντέλα και την εξέλιξη των δομών (schemas), πολύ λίγη προσοχή έχει δοθεί μέχρι τώρα στην εξέλιξη των εννοιών" [4] και άρα των οντοτήτων.

Σε πρόσφατες σχετικά δημοσιεύσεις επισημαίνεται το γεγονός ότι " μέχρι τώρα οι εννοιολογικές γλώσσες μοντελοποίησης - όπως η ER, UML, Description Logics στηρίζονται όλες στην

οντότητα, η οποία θεωρείται ότι παραμένει αμετάβλητη και "αυτοτελής" (atomic) ακόμα και όταν η κατάσταση της αλλάζει κατά την διάρκεια ζωής της. [4]

Το χαρακτηριστικό αυτό, δυστυχώς, αποτρέπει τις υφιστάμενες γλώσσες μοντελοποίησης από την καταγραφή των φαινομένων που αφορούν την εξέλιξη των υπαρχουσών οντοτήτων σε κάτι άλλο, όπως π.χ. Του ότι μια κάμπια γίνεται πεταλούδα, ή του ότι η Γερμανία αμέσως μετά τον Β' Παγκόσμιο Πόλεμο χωρίστηκε σε δύο άλλα κράτη.



Εικ. 2.1 Η εξέλιξη της Γερμανίας

Από "Flavio Rizzolo, Yannis Velegrakis, John Mylopoulos, Siarhei Bykau: Modeling Concept Evolution: A Historical Perspective. In ER, pages 331-345, 2009."

Γίνεται επομένως σαφές ότι μεγαλύτερη προσοχή πρέπει να δοθεί στην οντότητα, ως ένα αυτόνομο σύνολο ιδιοτήτων και των αντίστοιχων τιμών τους, χωρίς αυτό να σημαίνει ότι το σύνολο των ιδιοτήτων αυτών δεν μπορεί να αλλάξει με το χρόνο ή ότι η οντότητα η ίδια δεν μπορεί να πάψει να περιγράφεται από το συγκεκριμένο σύνολο. Σε αυτή την άποψη συμφωνούν και εργασίες που εξετάζουν το θέμα από την άποψη των δεδομενοχώρων (dataspaces)[2]. Πιο συγκεκριμένα, ο dataspace ορίζεται σαν μια αφαιρετική έννοια της διαχείρισης δεδομένων για ποικίλες εφαρμογές διαχείρισης κυβερνητικών και επιχειρησιακών δεδομένων, ψηφιακών βιβλιοθηκών, έξυπνων σπιτιών και προσωπικών πληροφοριών, ενώ η DataSpaceSupportPlatform ως ένα σύστημα που πρέπει να χτιστεί, για να παράσχει τις

απαιτούμενες από τους dataspaces υπηρεσίες. Σε αντίθεση με τα συστήματα ενσωμάτωσης δεδομένων, οι dataspaces δεν απαιτούν πλήρη σημασιολογική ενσωμάτωση των πόρων, προκειμένου να παράσχουν χρήσιμες υπηρεσίες, και προσεγγίζουν τα δεδομένα με μια λογική συνύπαρξης περισσότερο. Κύριος σκοπός τους είναι να παράσχουν μια βασική λειτουργικότητα επί όλων των πηγών δεδομένων ανεξάρτητα από το πόσο ενοποιημένες είναι οι πηγές αυτές. Για παράδειγμα μπορούν να προσφέρουν αναζήτηση με λέξεις – κλειδιά ανάλογη με αυτή της αναζήτησης επιφάνειας εργασίας.

Μια οντότητα μπορεί να αλλάξει με διάφορους τρόπους. [4]. Συγκεκριμένα μπορεί να έχουμε χρονική, μερεολογική και αιτιώδη εξέλιξη, όπου με τον όρο μερεολογική νοείται σχέση "μέρος του" (part of) μεταξύ των εννοιών, με τον όρο 'αιτιώδης εξέλιξη' νοείται σχέση "γίνεται" μεταξύ των εννοιών και εισάγεται και η έννοια του συνδέσμου, όπου ένας σύνδεσμος μεταξύ δύο εννοιών είναι τμήμα μιας τουλάχιστον από τις έννοιες με τις οποίες σχετίζεται και έχει κάποια αιτιώδη σχέση με ένα τμήμα της άλλης. Εισάγονται τέσσερις όροι εξέλιξης:

1. Ένταξη: π.χ. μια χώρα εντάσσεται στην Ευρωπαϊκή Ένωση
2. Διάσπαση: π.χ. η Γερμανία διασπάστηκε σε Ανατολική και Δυτική
3. Συγχώνευση: π.χ. η Τράπεζα Α συγχωνεύτηκε με την Τράπεζα Β
4. Απόσχιση: π.χ. η θυγατρική Εταιρεία αποσχίστηκε από την μητρική της

Προκύπτει λοιπόν η ανάγκη να ενταχθεί και η εξέλιξη της οντότητας στους παράγοντες επίδρασης της συμπεριφοράς της στο Web, άρα και η ανάγκη να μελετηθεί.

Κεφάλαιο 3
Μοντελοποίηση της εξέλιξης των Αντικειμένων

Στο κεφάλαιο αυτό θα ορίσουμε το πρόβλημα που μελετάμε, καθώς και τη μοντελοποίηση των δεδομένων. Πιο συγκεκριμένα θα παρουσιαστούν οι έννοιες "αντικείμενο", "ιδιότητες", "τιμές", "χαρακτηριστικά", "κατηγορία" (Κεφ. 3.1) και θα οριστεί η έννοια της εξέλιξης (Κεφ. 3.2).

3.1 Αντικείμενα, Ιδιότητες, Τιμές, Χαρακτηριστικά. Η έννοια της Κατηγορίας.

Οι οντότητες έχουν πλέον γίνει μια βασική δομή για την μοντελοποίηση των δεδομένων στις σύγχρονες εφαρμογές Ιστού. Χρησιμοποιούνται για να μοντελοποιήσουν τα δεδομένα που περιγράφουν τα αντικείμενα του πραγματικού κόσμου [2], όπως είναι οι άνθρωποι, οι χώρες και τα γεγονότα. Το πιο γνωστό παράδειγμα εφαρμογής ιστού που βασίζεται σε οντότητες είναι η Wikipedia, η οποία αποτελεί μια τεράστια “βιβλιοθήκη” που δημιουργήθηκε από τους χρήστες της ίδιας της εφαρμογής.

Η οντότητα είναι, λοιπόν, σχεδιαστικό κατασκεύασμα και χρησιμοποιείται, για να παραστήσει ένα αντικείμενο του πραγματικού κόσμου. Αυτός ο ορισμός είναι σύμφωνος με τους ορισμούς που ακολουθούνται και από άλλες σχετικές μελέτες [3].

Ορισμός 1:

Μια οντότητα e (entity) είναι ένα σύνολο ιδιοτήτων (properties) $P=\{p_1,p_2,...p_n\}$ που περικλείονται κάτω από ένα μοναδικό αναγνωριστικό (unique identifier) e_{id} που την προσδιορίζει απόλυτα. Γράφουμε $e_{id}:\{p_1,p_2,...p_n\}$ ή ισοδύναμα $e_{id}:P$

Μια ιδιότητα p (property) οντότητας είναι μια έννοια που περιγράφεται από ένα μοναδικό αναγνωριστικό p_{id} και περιγράφει κάποιο τμήμα της οντότητας.

<u>*Παράδειγμα:*</u> στην οντολογία της Wikipedia version 3.6 υπάρχει η οντότητα e Albert_Einstein με μοναδικό αναγνωριστικό <http://dbpedia.org/resource/Albert_Einstein> και ιδιότητες {academicadvisors, birthdate, caption, dateofbirth, dateofdeath, deathdate, deathplace, doctoraladvisor, ethnicity, fields, lccn, pnd, residence, shortdescription, signature, viaf)

Στη Χρονική Βάση Δεδομένων Οντοτήτων που δημιουργήθηκε -και περιγράφεται αναλυτικά στο κεφ. 4 - η οντότητα αυτή έχει μοναδικό αναγνωριστικό το $e_{id}=11$ και e_{id}:{p_{432}, p_{1390}, p_{1817}, p_{2924}, p_{2940}, p_{3001}, p_{3015}, p_{3250}, p_{3573}, p_{3840}, p_{5570}, p_{7752}, p_{8586}, p_{9536}, p_{9580}, p_{10820}}. Κάθε μέλος του συνόλου P είναι το μοναδικό αναγνωριστικό (p_{id}) της αντίστοιχης ιδιότητας που περιγράφει την οντότητα στην Wikipedia. Έτσι το p_{1390} αντιστοιχεί στην ιδιότητα birthdate.

Ορισμός 2:

Χαρακτηριστικό (attribute) a_i μιας οντότητας είναι ένα ζεύγος μίας ιδιότητας(property) p_i και μίας τιμής (value) v_i για την συγκεκριμένη ιδιότητα.

Γράφουμε: $a_i = (p_i , v_i)$

Τιμή (value) v_i μιας ιδιότητας p_i είναι κάθε τύπος δεδομένων, απλός (π.χ. string, integer, float) ή σύνθετος (π.χ list) ή και το αναγνωριστικό μιας άλλης οντότητας.

<u>*Παράδειγμα:*</u> η οντότητα Albert_Einstein στην ιδιότητα ethnicity έχει τιμή Jewish.

Ορισμός 3:

Στιγμιότυπο (instance) i μιας οντότητας e είναι ένα σύνολο χαρακτηριστικών της (attributes) $\{a_1,a_2,...,a_n\}$, με ίδιο πληθάριθμο με το σύνολο P των ιδιοτήτων της, σε συγκεκριμένο και γνωστό χρόνο αναφοράς. Γράφουμε $i(e)= \{a_1,a_2,...,a_n\}$ ή ισοδύναμα $i(e)= \{(p_1, v_1), (p_2, v_2), ..., (p_i , v_i)\}$

Ιδιότητες – μέλη μπορούν να αφαιρούνται από το σύνολο P των ιδιοτήτων μιας οντότητας. Το σύνολο P είναι σε κάθε χρονική στιγμή τουλάχιστον γνήσιο υποσύνολο του συνόλου των

ιδιοτήτων που είχε, έχει ή θα έχει ποτέ η οντότητα. Στην βιβλιογραφία συναντούμε παρόμοιους ορισμούς [3] και όπως αναφέρεται χαρακτηριστικά "Στην πραγματικότητα το σύνολο των γνωρισμάτων (σ.σ. Ιδιοτήτων) για τα οποία έχουμε τιμές μπορεί να εξελίσσεται (βλ. Εικόνα 3.1), καθώς επίσης εξελισσόμενο μπορεί να είναι και το ίδιο των σύνολο των γνωρισμάτων (σ.σ. ιδιοτήτων) καθ' εαυτό ." [1] (βλ. Εικόνα 3.2)

Παράδειγμα: στην Εικόνα 3.1 βλέπουμε ότι για την οντότητα John Malkovitch έχουμε διαφορετικό πλήθος ιδιοτήτων σε δύο διαφορετικές Wikipedia εκδόσεις 3.5.1(3/2010) και 3.8(6/2012), ενώ στην Εικ. 3.2 βλέπουμε αλλαγές στις τιμές των ιδιοτήτων "alt" & "placeofbirth" για τον Brad Pitt.

Ανοικτό Πανεπιστήμιο Κύπρου
Open University of Cyprus

PES700:Routzouni,E.: Observing Objects Evolution on the Web

Scientific Coordinator: Thanasis Hadzilakos

Supervisor Professor : Aikaterini Ioannou

Which are the properties of John Malkovich in each one of the Wikipedia v3.5.1,Wikipedia v3.8?

Applications that study and incorporate data form various sources in order to correlate entities' instances could benefit form the results of this queery

Rows Selected:22

Ιδιότητα	στην Wikipedia v3.5.1	στην Wikipedia v3.8
almamater	ΑΝΥΠΑΡΚΤΗ/NON-EXISTANT	ΥΠΑΡΧΕΙ/EXISTS
awards	ΑΝΥΠΑΡΚΤΗ/NON-EXISTANT	ΥΠΑΡΧΕΙ/EXISTS
children	ΑΝΥΠΑΡΚΤΗ/NON-EXISTANT	ΥΠΑΡΧΕΙ/EXISTS
education	ΑΝΥΠΑΡΚΤΗ/NON-EXISTANT	ΥΠΑΡΧΕΙ/EXISTS
hometown	ΑΝΥΠΑΡΚΤΗ/NON-EXISTANT	ΥΠΑΡΧΕΙ/EXISTS
nationality	ΑΝΥΠΑΡΚΤΗ/NON-EXISTANT	ΥΠΑΡΧΕΙ/EXISTS

Εικόνα 3.1 Πλήθος ιδιοτήτων της οντότητας John Malkovitch στις Wikipedia versions 3.5.1 & 3.8.

Ανοικτό Πανεπιστήμιο Κύπρου
Open University of Cyprus

PES700:Routzouni,E.: Observing Objects Evolution on the Web

Scientific Coordinator: Thanasis Hadzilakos

Supervisor Professor : Aikaterini Ioannou

Which properties of Brad_Pitt, existing in both Wikipedia v3.5.1,Wikipedia v3.6 and having only one value in each one of them , differ in this value?

Applications that study and incorporate data form various sources in order to correlate entites' instances could benefit form the results of this queery

Rows Selected:2

Ιδιότητα	Τιμή Wikipedia v3.5.1	Τιμη Wikipedia v3.6
alt	A Caucasian male in his mid-40s with brown hair. He is wearing a black suit and white shirt with a black bow-tie.	A Caucasian with light brown hair, blue eyes and a short brown beard, in front of a turquoise background. He is wearing a white shirt and white hat.
placeofbirth	U.S.	Shawnee, Oklahoma, U.S.

Εικόνα 3.2 Αλλαγές στις τιμές των ιδιοτήτων "alt" και "placeofbirth" για την οντότητα Brad Pitt

Ομοίως και διαφορές στον πληθάριθμο $N_{i(e)}$ Ne:P σηματοδοτούν εξελικτικές σχέσεις και σε παρόμοια θεώρηση φαίνεται να συμφωνούν και άλλοι ερευνητές [4].

Η θεώρηση αυτή είναι διαφορετική από την ακόλουθη:

«Συγκεκριμένα πολλά στιγμιότυπα μιας έννοιας είναι πιθανό να έχουν τιμές καθορισμένες από ένα σύνολο γνωρισμάτων (σ.σ εδώ νοείται ιδιότητα σύμφωνα με τους δικούς μας ορισμούς), αλλά δεν σημαίνει ότι όλα τα στιγμιότυπα της έννοιας θα έχουν τιμές για όλα τους τα γνωρίσματα».[1] Εμείς επεκτείνουμε τους υφιστάμενους παρόμοιους ορισμούς, ώστε να μην μπορούμε να αποδώσουμε ένα στιγμιότυπο σε μια έννοια/οντότητα, αν δε γνωρίζουμε όλες "τις τιμές των γνωρισμάτων του" [1], ή ισοδύναμα, κατά τους δικούς μας ορισμούς, όλα "τα χαρακτηριστικά του".

Αυτό εύκολα φαίνεται από το ακόλουθο παράδειγμα:

Έστω οι οντότητες Human και Comic Character με σύνολα ιδιοτήτων

P_1={name, surname, townOfresidence, job,color, canFly) και

P_2={name, townOfresidence, job, color, canFly) αντίστοιχα .

Η ιδιότητα canFly είναι τύπου string, ενώ η ιδιότητα color είναι τύπου boolean.

Έστω τώρα ότι ένα στιγμιότυπο με χαρακτηριστικά (p_i,v_i) έχει μεταξύ άλλων τα χαρακτηριστικά : (χρώμα, πράσινο) και (ιπτάμενο, true). Σε ποια οντότητα θα αποδοθεί;

Προφανώς στην δεύτερη, αφού δεν υπάρχουν ιπτάμενοι πράσινοι άνθρωποι. Αν όμως τα χαρακτηριστικά ήταν του τύπου (χρώμα, λευκό) και (ιπτάμενο, false), τότε το στιγμιότυπο θα μπορούσε να αποδοθεί και στην πρώτη οντότητα.

Ένα άλλο ερώτημα που αναδύεται από τα παραπάνω είναι και το εξής: ποιο είναι το ελάχιστο σύνολο γνωρισμάτων (δηλ. ιδιοτήτων με τις αντίστοιχες τιμές τους) τα οποία αν έχει ένα αντικείμενο μπορεί να αποδοθεί σε μια οντότητα;

Πιστεύουμε πως, εάν θέλουμε να κινηθούμε σε Web εννοιών, θα πρέπει να μπορούμε να θέτουμε σχετικά σαφή όρια στο πού ξεκινά μια οντότητα και που σταματά μια άλλη. Το παραπάνω ερώτημα επιδέχεται πιθανολογικής απάντησης, μέσω διαδικασίας εκπαιδευόμενης μάθησης με δεδομένα εκπαίδευσης όλα τα γνωστά στιγμιότυπα για μια οντότητα, η οποία ξεφεύγει από τα όρια της παρούσας μεταπτυχιακής διατριβής.

Συνοψίζοντας:

- η οντότητα μοντελοποιεί ένα μόνο αντικείμενο του πραγματικού κόσμου
- ένα αντικείμενο μέσα στο χρόνο μοντελοποιείται από πολλές διαφορετικές οντότητες (στιγμιότυπο αντικειμένου = οντότητα)
- κάθε οντότητα διακρίνεται από το μοναδικό της αναγνωριστικό e.

Η παρούσα μεταπτυχιακή διατριβή θα ασχοληθεί με οντότητες που μοντελοποιούν πρόσωπα και συγκεκριμένα με τις οντότητες-πρόσωπα που εμφανίζονται στις διάφορες εκδόσεις της Wikipedia, καθώς αυτές αναθεωρούνται με την πάροδο του χρόνου.

Σε αυτό το σημείο θα εισάγουμε και την έννοια της κατηγορίας για τα πρόσωπα που μελετήθηκαν:

Μια κατηγορία είναι μια ομαδοποίηση προσώπων με βάση κάποιο κοινό χαρακτηριστικό τους, για το οποίο όμως έγιναν ευρύτερα γνωστά (και το οποίο είχε ως αποτέλεσμα να συμπεριληφθούν στα λήμματα της Wikipedia).

Για παράδειγμα:

Ο Φιλόσοφος Αριστοτέλης έχει ένα σύνολο από γνωρίσματα π.χ. ημερομηνία γέννησης, ημερομηνία θανάτου, τόπο γέννησης, πρόσωπα τα οποία επηρέασε, όνομα. Η κατηγορία στην οποία εντάσσεται όμως είναι αυτή του φιλοσόφου και όχι του ηθοποιού (όπως ο George Clooney) ή του μουσικού καλλιτέχνη (όπως ο George Michael), που επίσης έχουν ένα παρόμοιο σύνολο ιδιοτήτων (π.χ. ημερομηνία γέννησης, ημερομηνία θανάτου, τόπο γέννησης,..,), διότι για αυτή του την δραστηριότητα είναι ευρύτερα γνωστός.

3.2 Ορισμός της εξέλιξης των Αντικειμένων.

Με τον όρο "Εξέλιξη των Αντικειμένων" στην παρούσα μεταπτυχιακή διατριβή θα εννοούμε τις αλλαγές που συμβαίνουν:

- στο πλήθος του συνόλου P των ιδιοτήτων,
- στα χαρακτηριστικά a_i ,
- στην πληθυκότητα των χαρακτηριστικών

αντικειμένων - προσώπων που εμφανίζονται στην Wikipedia (που χρησιμοποιείται ως παράδειγμα εφαρμογής ιστού με ευμετάβλητα δεδομένα) σε διαφορετικές χρονικές περιόδους που οριοθετούνται από τις αναθεωρήσεις των εκδόσεών της.

Με τον όρο πληθυκότητα εννοούμε πόσες φορές μπορεί να απαντά μια ιδιότητα p μιας οντότητας e ως πρώτο μέλος ενός ζεύγους $a_κ=(p_κ,v_κ)$ που ανήκει στο σύνολο i(e) των χαρακτηριστικών της οντότητας.

Για παράδειγμα για το πρόσωπο Aristotle η ιδιότητα "influences" (δέχτηκε επιρροές) έχει πολλές τιμές : Democritus, Heraclitus, Parmenides, Plato, Socrates.

Έτσι για το υπόλοιπο του κειμένου:

- οι έννοιες αντικείμενο (Object), οντότητα (entity), αναγνωριστικό οντότητας (entity identifier), στιγμιότυπο (instance), πρόσωπο (person), person identifier (pid), καθώς και
- οι έννοιες ιδιότητα (property) , αναγνωριστικό ιδιότητας (property identifier),

θα αναφέρονται και θα χρησιμοποιούνται ισοδύναμα.

- Η έννοια χαρακτηριστικό θα αποδίδει, κατά τα προαναφερθέντα, μια ιδιότητα με την τιμή της.

Παραδείγματα:

Ο Αριστοτέλης Ωνάσης στη Wikipedia v. 3.5.1 εμφανίζεται να περιγράφεται από 16 ιδιότητες, κατά 4 λιγότερες από αυτές που τον περιγράφουν στην αμέσως επόμενη της χρονικά v.3.6 (Εικ. 3.3). Ο Bill Clinton περιγράφεται με μόλις μια ιδιότητα παραπάνω στην v.3.6 (31) έναντι της v. 3.5.1 (30) (Εικ. 3.4), αλλά 4 από τις ιδιότητές του έχουν αλλάξει τιμή στη νεότερη v. 3.6. (Εικ.3.5)

Παρατηρήστε ότι τα παραπάνω ερωτήματα εστιάζουν σε ποσοτικές μεταβολές. Δεν είναι όμως μόνο αυτές οι μεταβολές που μας ενδιαφέρουν. Θα θέλαμε να γνωρίζουμε ποιο πρόσωπο ή καλύτερα ποια/ποιες κατηγορία/ες προσώπων έχουν τις περισσότερες ιδιότητες. Ή ποιες ιδιότητες αλλάζουν τιμές πιο πολύ. Ή ποιες ιδιότητες προστέθηκαν (αφαιρέθηκαν) από την μετάβαση από μια έκδοση σε νεότερη. Ή ποιες είναι οι κατηγορίες με τα περισσότερα μέλη.

Για παράδειγμα: Η κατηγορία "μουσικοί " περιγράφονται από 440 επιπλέον ιδιότητες στην Wikipedia έκδοση 3.8.0, οι "συγγραφείς" από 235, οι "ηθοποιοί" από 197, αλλά τα σκήπτρα κρατούν οι κάτοχοι κυβερνητικών θέσεων (office holders), που αύξησαν κατά 461 τις ιδιότητες που απαντούν σε πρόσωπα της κατηγορίας τους.

Η πολυπληθέστερη κατηγορία στην Wikipedia v 3.5.1 είναι οι μουσικοί καλλιτέχνες.

Ανοικτό Πανεπιστήμιο Κύπρου
Open University of Cyprus

ΠΕΣ700:Ρουτζούνη,Ε.: Παρακολούθηση της εξέλιξης των Αντικειμένων στο Διαδίκτυο

Επιστημονικός Υπεύθυνος: Θανάσης Χατζηλάκος

Επιβλέπουσα Καθηγήτρια : Αικατερίνη Ιωάννου

Ποιες ιδιότητες έχει ο/η Aristotle Onassis σε κάθε μια από τις Wikipedia v3.5.1,Wikipedia v3.6;
Εφαρμογές που μελετούν και ενσωματώνουν δεδομένα από διαφορετικές πηγές και συσχετίζουν στιγμιότυπα οντοτήτων μεταξύ τους θα αξιοποιούσαν τα αποτελέσματα του ερωτήματος αυτού.
Αριθμός εγγραφών που ανασύρθηκαν:24

Ιδιότητα	στην Wikipedia v3.5.1	στην Wikipedia v3.6
dateofbirth	ΑΝΥΠΑΡΚΤΗ/NON-EXISTANT	ΥΠΑΡΧΕΙ/EXISTS
dateofdeath	ΑΝΥΠΑΡΚΤΗ/NON-EXISTANT	ΥΠΑΡΧΕΙ/EXISTS
placeofbirth	ΑΝΥΠΑΡΚΤΗ/NON-EXISTANT	ΥΠΑΡΧΕΙ/EXISTS
placeofdeath	ΑΝΥΠΑΡΚΤΗ/NON-EXISTANT	ΥΠΑΡΧΕΙ/EXISTS
below	ΥΠΑΡΧΕΙ/EXISTS	ΥΠΑΡΧΕΙ/EXISTS

Εικόνα 3.3 Εξέλιξη των ιδιοτήτων της οντότητας Αριστοτέλης Ωνάσης σε 2 Wikipedia versions (3.5.1 & 3.6)

Ανοικτό Πανεπιστήμιο Κύπρου
Open University of Cyprus

ΠΕΣ700:Ρουτζούνη,Ε.: Παρακολούθηση της εξέλιξης των Αντικειμένων στο Διαδίκτυο

Επιστημονικός Υπεύθυνος: Θανάσης Χατζηλάκος

Επιβλέπουσα Καθηγήτρια : Αικατερίνη Ιωάννου

Πόσες ιδιότητες έχει ο/η Bill Clinton στην Wikipedia v3.5.1 και πόσες στην Wikipedia v3.6;
Εφαρμογές που μελετούν και ενσωματώνουν δεδομένα από διαφορετικές πηγές και συσχετίζουν στιγμιότυπα οντοτήτων μεταξύ τους θα αξιοποιούσαν τα αποτελέσματα του ερωτήματος αυτού.
Αριθμός εγγραφών που ανασύρθηκαν:1

Πλήθος Ιδιοτήτων στην Wikipedia v3.5.1	Πλήθος Ιδιοτήτων στην Wikipedia v3.6
30	31

Εικόνα 3.4 Εξέλιξη του πλήθους ιδιοτήτων της οντότητας Bill_Clinton σε 2 Wikipedia versions (3.5.1 & 3.6)

Ανοικτό Πανεπιστήμιο Κύπρου
Open University of Cyprus

ΠΕΣ700:Ρουτζούνη,Ε.: Παρακολούθηση της εξέλιξης των Αντικειμένων στο Διαδίκτυο

Επιστημονικός Υπεύθυνος: Θανάσης Χατζηλάκος

Επιβλέπουσα Καθηγήτρια : Αικατερίνη Ιωάννου

Ποιές ιδιότητες του/της Bill Clinton, μονότιμες και κοινές στις Wikipedia v3.5.1 και Wikipedia v3.6, έχουν διαφορετική τιμή στη δεύτερη ;
Εφαρμογές που μελετούν και ενσωματώνουν δεδομένα από διαφορετικές πηγές και συσχετίζουν στιγμιότυπα οντοτήτων μεταξύ τους θα αξιοποιούσαν τα αποτελέσματα του ερωτήματος αυτού.
Αριθμός εγγραφών που ανασύρθηκαν:4

Ιδιότητα	Τιμή Wikipedia v3.5.1	Τιμη Wikipedia v3.6
birthplace	Hope%2C_Arkansas	Hope, Arkansas
party	Democratic_Party_%28United_States%29	Democratic
placeofbirth	Hope%2C_Arkansas	Hope, Arkansas
website		William J. Clinton Presidential Library

Σελίδα

Εικόνα 3.5 Αλλαγές στις τιμές των κοινών ιδιοτήτων στις versions 3.5.1 & 3.6 για τον Bill_Clinton

Κεφάλαιο 4
Μεθοδολογία

Στο κεφάλαιο αυτό θα περιγραφεί το είδος των δεδομένων που μελετήθηκαν στην παρούσα μεταπτυχιακή διατριβή, ο λόγος για τον οποίο επιλέχθηκαν, το σκοπούμενο αποτέλεσμα (4.1), καθώς και η μεθοδολογία που ακολουθήθηκε και πιο συγκεκριμένα: Στο 4.2 θα περιγραφεί η μεθοδολογία εξόρυξης των προσώπων και των χαρακτηριστικών τους, στο 4.3 η επεξεργασία που έλαβε χώρα σε δεδομένα που αφορούν πρόσωπα, στο 4.4 τα ερευνητικά ερωτήματα που επιλέχθηκαν να τεθούν, στο 4.5 η μέθοδος υλοποίησης των ερευνητικών ερωτημάτων και στο 4.6 ο τρόπος που επιλέχτηκε για την εύκολη ενσωμάτωση custom επιλογών στο σύστημα που αναπτύχθηκε, το οποίο περιγράφεται αναλυτικά στο κεφ. 5 του παρόντος.

4.1 Τα δεδομένα που μελετήθηκαν

Στα πλαίσια της επίτευξης του στόχου της παρούσας μεταπτυχιακής διατριβής, όπως αυτός περιγράφηκε στο κεφ. 1.2 του παρόντος, χρησιμοποιήθηκαν δεδομένα από την Αγγλική έκδοση της Wikipedia.

Η Wikipedia (σύνθετη λέξη από το Χαβανέζικο wiki, που σημαίνει «γρήγορος» και το ελληνικής προέλευσης encyclopedia που σημαίνει εγκυκλοπαίδεια) είναι μία πολυγλωσσική διαδικτυακή εγκυκλοπαίδεια. Τα λήμματά της παρέχουν συνδέσμους που οδηγούν το χρήστη σε σχετικές

σελίδες με επιπλέον πληροφορίες. Η Wikipedia γράφεται με συνεργασία εθελοντών από όλο τον κόσμο και οποιοσδήποτε χρήστης μπορεί να επεξεργαστεί λήμματα, πατώντας στο σύνδεσμο "επεξεργασία" που εμφανίζεται στην κορυφή της κάθε σελίδας της. Από την ίδρυσή της, το 2001, η Wikipedia έχει μεγαλώσει ραγδαία και έχει εξελιχθεί σε έναν από τους μεγαλύτερους ιστότοπους αναφοράς στο Web, ενώ μόνο το έτος 2008 προσέλκυσε τουλάχιστον 684 εκατομμύρια επισκέπτες. Το Μάρτιο του 2013 (οπότε και αντλούνται τα αναφερόμενα στοιχεία) υπάρχουν πάνω από 100.000 ενεργοί συνεισφέροντες που εργάζονται σε περισσότερα από 25.000.000 άρθρα και σε περισσότερες από 285 γλώσσες. Μέχρι σήμερα υπάρχουν 4,1 εκατομμύρια λήμματα μόνο στην Αγγλική Wikipedia.

(πηγή: http://en.wikipedia.org/wiki/Wikipedia, απόδοση στα ελληνικά Ε.Ρουτζούνη)

Κάθε λήμμα της Wikipedia είναι και ένα αντικείμενο. Καθημερινά εκατοντάδες χιλιάδες επισκέπτες από όλο τον κόσμο κάνουν δεκάδες χιλιάδες αλλαγές και δημιουργούν χιλιάδες νέα λήμματα για την ενίσχυση της γνώσης που βρίσκεται σε αυτήν. ***Έτσι με την πάροδο του χρόνου νέα αντικείμενα προστίθενται στην Wikipedia και τα παλαιότερα ενδεχόμενα εμπλουτίζονται με νέες πληροφορίες σχετικές με αυτά.***

Οι αλλαγές αυτές απεικονίζονται στις διάφορες αναθεωρήσεις των εκδόσεων της Wikipedia (versions) και σήμερα βρισκόμαστε στην version 3.8.

Πληροφορίες για την οντολογία της Wikipedia είναι διαθέσιμες στην URL http://mappings.dbpedia.org/server/ontology/classes/,

ενώ στην URL http://wiki.dbpedia.org/Downloads38 υπάρχουν επίσης διαθέσιμα, για download σε διάφορες γλώσσες, τα αρχεία (Raw Infobox Properties.en.nt.bz2) που περιέχουν τα δεδομένα (αντικείμενα – λήμματα) της τρέχουσας έκδοσης που, την περίοδο της εκπόνησης της παρούσας μεταπτυχιακής διατριβής, (Οκτώβριος 2012 - Μάιος 2013) είναι η v.3.8.

Έτσι, κατά την άποψή μας, η Wikipedia αποτελεί ένα περιβάλλον πολύ κατάλληλο για να μελετηθούν φαινόμενα εξέλιξης οντοτήτων και αυτό γατί διαθέτει:

1. συγκεκριμένη οντολογία, η οποία εύκολα μπορεί να ανακτηθεί από τον σύνδεσμο http://mappings.dbpedia.org/server/ontology/classes/

2. δεδομένα εύκολα προσβάσιμα προς επεξεργασία με σύνδεσμο εκκίνησης τον http://wiki.dbpedia.org/Downloads38

3. μεγάλο πλήθος δεδομένων στην αγγλική της έκδοση που σημαίνει ότι μπορούν από την μελέτη τους να προκύψουν στατιστικά ασφαλή συμπεράσματα

4. σαφή χρονικά ορόσημα (τουλάχιστον 9) για μελέτη της εξέλιξης των οντοτήτων στο χρόνο, που είναι τόσα όσες και οι διαθέσιμες αναθεωρημένες εκδόσεις (versions) της Wikipedia

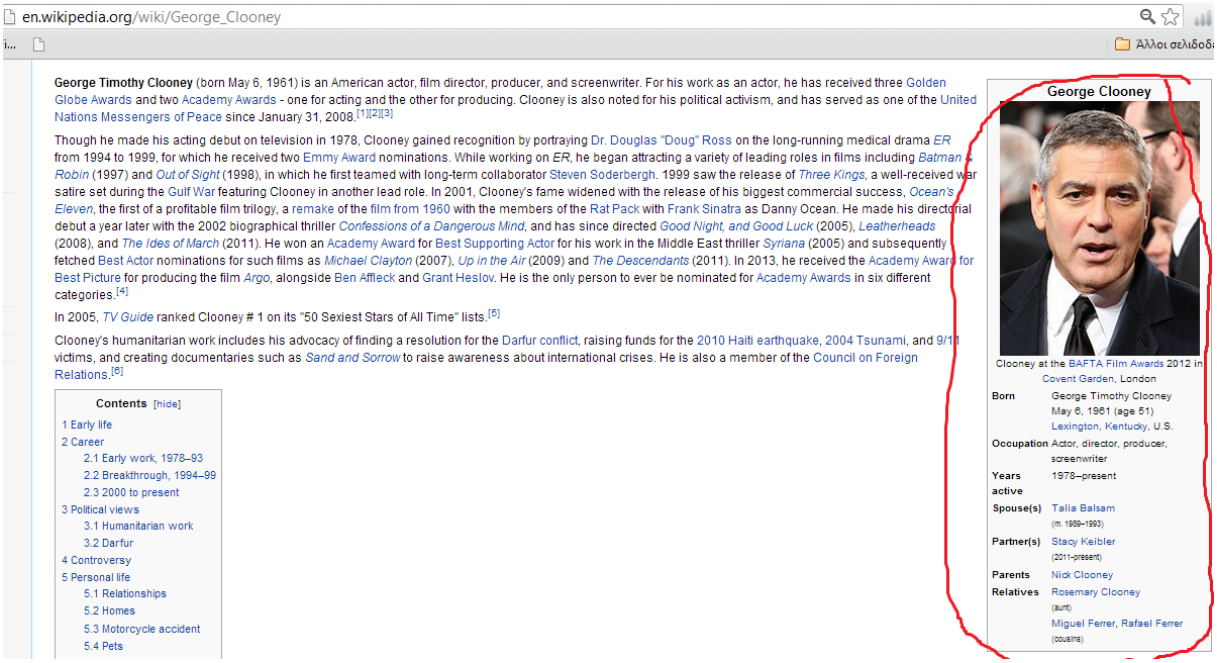
και καλύπτουν ένα διάστημα από τον Ιούλιο του 2007 έως και τον Ιούνιο του 2012, δηλαδή μια σχεδόν πλήρη πενταετία.

5. ***ευμετάβλητα δεδομένα***, αφού δημιουργείται από χρήστες ***και επομένως εμφανίζει τα σχετικά προβλήματα*** π.χ entity matching που εμφανίζεται, όταν πολλές οντότητες μοντελοποιούν το ίδιο αντικείμενο του πραγματικού κόσμου ή ποια πληροφορία του ιστού υπεισέρχεται σε μια συγκεκριμένη οντότητα και ή σε ένα συγκεκριμένο στιγμιότυπο μιας οντότητας [2,3] και

6. ***ένα μοναδικό αναγνωριστικό για κάθε εμφανιζόμενο σε αυτήν αντικείμενο,*** το οποίο μπορεί να χρησιμοποιηθεί για να εντοπίσουμε το αντικείμενο αυτό στις διαφορετικές εκδόσεις.

Αποφασίστηκε να μελετηθεί η εξέλιξη στο χρόνο μόνο για οντότητες που μοντελοποιούν προσώπα. Αυτό έγινε για λόγους:

- απλότητας: η μεθοδολογία παραμένει ίδια είτε ασχολούμαστε με πρόσωπα είτε με το σύνολο των αντικειμένων, αλλάζουν όμως πολύ οι απαιτούμενοι για την επεξεργασία τους χρόνοι και χώροι και
- σκοπιμότητας: θεωρήθηκε ότι οι αλλαγές σε ιδιότητες και πληροφορίες που αφορούν πρόσωπα είναι πιο δυναμικές.



Εικ. 4.1 Η σελίδα της Αγγλικής Wikipedia για τον George Clooney. Δεξιά με κόκκινο κύκλο το info-box.

Μελετήθηκαν δεδομένα από τις εκδόσεις (versions) 3.5.1, την 3.6 και την 3.8.

Τα Wikipedia dumps έλαβαν χώρα για την έκδοση 3.5.1 τέλη Μαρτίου του 2010, για την έκδοση 3.6 τον Νοέμβριο του 2011 και για την έκδοση 3.8 (τρέχουσα) τέλη Ιουνίου του 2012. Οι εκδόσεις 3.5.1 και 3.8 λοιπόν απέχουν μεταξύ τους χρονικά πάνω από 24 μήνες (27), οι εκδόσεις 3.5.1 και 3.6 απέχουν πάνω από 12 μήνες και λιγότερο από 24 (21), ενώ οι εκδόσεις 3.6 και 3.8 απέχουν λιγότερο από 12 μήνες (8). Η κατανομή αυτή των χρονικών διαστημάτων κρίθηκε ικανοποιητική για τους σκοπούς της μελέτης.

Στόχος μας, όπως περιγράφηκε στο κεφ. 1.2 του παρόντος, ήταν να εξορυχθούν, μετά από επεξεργασία των αρχείων αυτών, οι οντότητες που μοντελοποιούν πρόσωπα, καθώς και τα στιγμιότυπά τους (σύνολο χαρακτηριστικών, δηλ. σύνολο ζευγών ιδιότητας-τιμής, κατά τα οριζόμενα στο κεφ.3.1 του παρόντος), στους χρόνους αναφοράς, που αντιστοιχούν στις εκδόσεις που μελετήθηκαν και να ενταχθούν σε μια ενιαία "χρονικά ευαίσθητη" βάση δεδομένων οντοτήτων, στην οποία θα μπορούσαμε στην συνέχεια να θέτουμε κατάλληλα ερευνητικά ερωτήματα και να λαμβάνουμε αποτελέσματα, αποκαλυπτικά για την εξέλιξη των οντοτήτων-προσώπων στις υπό μελέτη χρονικές περιόδους

Τα ερωτήματα αυτά θα μπορούσαν να είναι τόσο ποιοτικά (π.χ ποιου είδους αντικείμενα εξελίσσονται συχνότερα) όσο και ποσοτικά (π.χ πόσες ιδιότητες προστίθενται σε ένα αντικείμενο) και πάντως τέτοια που, κατά την άποψή μας, θα συνεισέφεραν στην ενεργή έρευνα στο πεδίο. Παρακάτω (βλ. κεφ.4.4) θα ασχοληθούμε αναλυτικότερα με αυτά τα ερευνητικά ερωτήματα και τη σημασία τους.

4.2 Εξόρυξη των Προσώπων και των χαρακτηριστικών τους.

4.2.1 Παρατηρήσεις χρήσιμες για την εξόρυξη οντοτήτων-προσώπων.

Στην διεύθυνση http://wiki.dbpedia.org/Downloads38, όπως έχει ήδη αναφερθεί στο 4.1, βρίσκονται, διαθέσιμα προς download, τα αρχεία που προέκυψαν από το dump έκδοσης 3.8 (τρέχουσας κατά την περίοδο της εκπόνησης της μεταπτυχιακής διατριβής αυτής). Πρόκειται για το αρχείο Raw infobox_properties_en.nt.bz2 και υπάρχει ένα τέτοιο αρχείο για κάθε μια από τις κύριες γλώσσες στις οποίες δημοσιεύτηκε η συγκεκριμένη έκδοση της Wikipedia. Επίσης από τη διεύθυνση αυτή εκκινούν σύνδεσμοι για τις παλαιότερες Wikipedia εκδόσεις, που και αυτές διαθέτουν αντίστοιχα αρχεία..

Κάθε ένα από αυτά τα αρχεία περιέχει γραμμές με την εξής μορφή:

<http://dbpedia.org/resource/Aristotle> <http://dbpedia.org/property/region> "Western philosophy"@en . (μια γραμμή-1)

<http://dbpedia.org/resource/Aristotle> <http://dbpedia.org/property/era> <http://dbpedia.org/resource/Ancient_philosophy> .
(άλλη γραμμή-2)

<http://dbpedia.org/resource/Aristotle> <http://dbpedia.org/property/color> "#B0C4DE"@en .
(άλλη γραμμή-3)

Είναι φανερό ότι οι γραμμές εμφανίζουν συγκεκριμένη δομή.

Το tag *"<http://dbpedia.org/resource/.....>"* ακολουθεί το μοναδικό όνομα της οντότητας και άρα αποδίδει την οντότητα κατά τον Ορισμό 1 του κεφ.3.1 του παρόντος.

Το tag *"<http://dbpedia.org/property/.....>"*. ακολουθεί το όνομα της ιδιότητας που περιγράφει την οντότητα και άρα αποδίδει την ιδιότητα οντότητας κατά τον ίδιο Ορισμό.

Ακολουθεί η τιμή που μπορεί να είναι είτε κάποιος γνωστός τύπος δεδομένων είτε σύνδεσμος προς κάποια άλλη οντότητα της Wikipedia. (Ορισμός 2 του κεφ.3.1 του παρόντος.)

Πιο πάνω παρατέθηκε ένα τμήμα των εγγραφών για τον Αριστοτέλη. Ας παρατηρήσουμε τη γραμμή 2. Τη δομή "...resource/" ακολουθεί το "Aristotle". Επομένως κάθε γραμμή που θα περιγράφει τον Αριστοτέλη θα πρέπει να ξεκινά πάντα με το tag:
<http://dbpedia.org/resource/Aristotle>.

Η γραμμή 2, λοιπόν, αναφέρεται στην ιδιότητα με όνομα "era" της οντότητας Αριστοτέλης, αφού αυτή ακολουθεί το tag <http://dbpedia.org/property/...>, με τιμή για την ιδιότητα αυτή έναν άλλο σύνδεσμο Wikipedia, *τον <http://dbpedia.org/resource/Ancient_philosophy>*.

Βλέπουμε λοιπόν πως κάθε μια από τις γραμμές των αρχείων αυτών αποδίδει εν τέλει ένα "χαρακτηριστικό" της οντότητας στην οποία αναφέρεται, κατά τον Ορισμό 2 του κεφ.3.1 του παρόντος.

Σε κάθε τέτοιο αρχείο, το σύνολο των γραμμών για μια συγκεκριμένη οντότητα-πρόσωπο αποδίδει το στιγμιότυπο της οντότητας αυτής στη συγκεκριμένη Wikipedia έκδοση στην οποία το αρχείο αντιστοιχεί, κατά τον Ορισμό 3 του κεφ.3.1 του παρόντος.

Παρατηρήσαμε επίσης ότι, όταν τα δεδομένα των Raw Infoboxe Properties αρχείων αφορούν πρόσωπο, η ιδιότητα **<http://dbpedia.org/property/wikiPageUsesTemplate>** λαμβάνει κάποια από τις ακόλουθες τιμές:

- είτε τιμή της μορφής <http://dbpedia.org/resource/Template:Persondata>

- είτε τιμή της μορφής <http://dbpedia.org/resource/Template: Infobox_XXXXXX>

Σημειώνεται ότι δεν ξέρουμε ποιας μορφής τιμή θα συναντήσουμε ούτε σε ποια θέση ανάμεσα στο σύνολο των γραμμών που αφορούν σε ένα συγκεκριμένο πρόσωπο, αλλά ότι σίγουρα θα συναντήσουμε μια από τις δύο, μπορεί και τις δύο, αν τα δεδομένα μας αφορούν σε πρόσωπο.
Η πρώτη γραμμή μάς παρέχει απλώς την πληροφορία ότι τα δεδομένα μας αφορούν πρόσωπο.
Η δεύτερη γραμμή μάς παρέχει πληροφορία για το σε ποια κλάση της οντολογίας της Wikipedia ανήκει το συγκεκριμένο αντικείμενο (πρόσωπο) .
Η οντολογία της Wikipedia περιγράφεται, όπως έχουμε ήδη αναφέρει, στον σύνδεσμο:
http://mappings.dbpedia.org/server/ontology/classes/
Εκεί συναντάμε μεταξύ άλλων και την κλάση " Person ", που μοντελοποιεί πρόσωπα, με όλες τις θυγατρικές υποκλάσεις της.
Έτσι, όταν στην ιδιότητα "wikiPageUsesTemplate" αποδίδεται τιμή με χρήση του tag *<http://dbpedia.org/resource/Template: Infobox_XXXXXX>* , τότε το XXXXXX είναι πάντα κάποια από αυτές ακριβώς τις θυγατρικές κλάσεις της " Person". Με άλλα λόγια δηλώνεται κατευθείαν σε ποια υποκλάση ανήκει το πρόσωπο. Η τιμή της ιδιότητας αυτής αποδίδει την "κατηγορία" κατά τα οριζόμενα στο κεφ. 3.1**.**
Για το παράδειγμα του Αριστοτέλη, που χρησιμοποίηθηκε πιο πάνω, η τιμή της ιδιότητας <http://dbpedia.org/property/wikiPageUsesTemplate> είναι:
<http://dbpedia.org/resource/Template:Infobox_Philosopher>. Ο Αριστοτέλης ανήκει, επομένως, στην κατηγορία " Philosopher".

4.2.2 Υλοποίηση της εξόρυξης των οντοτήτων-προσώπων.

Αξιοποιώντας τις παραπάνω παρατηρήσεις (βλ.κεφ. 4.2.1) έγινε download, αποσυμπίεση και επεξεργασία με κώδικα Python v3.3 των Raw Infobox Properties.en.nt.bz2 αρχείων για τις versions 3.5.1, 3.6 και 3.8 της Αγγλικής Wikipedia.
Αποτέλεσμα της επεξεργασίας αυτής ήταν η δημιουργία μιας SQLite3 Βάσης Δεδομένων (ΒΔ), με μοναδικό περιεχόμενο αρχικά έναν πίνακα TEMP_persons568 (TEMPORARY) με στήλες (Όνομα,Κατηγορία) και πρωτεύον κλειδί το όνομα. Στον πίνακα αυτόν εξορύχθηκαν όλες οι οντότητες-πρόσωπα, καθώς και οι αντίστοιχες κατηγορίες τους από όλες τις επιλεγμένες προς μελέτη Wikipedia εκδόσεις**.**
Η υλοποίηση αυτή (κώδικας Python v.3.3 findpersons.py) αποτελεί τμήμα του component Persons' Data miner που περιγράφεται στο κεφ. 5.1

4.2.3 Δημιουργία των προσωρινών αρχείων με τα Πρόσωπα για κάθε Wikipedia version

Αφού τα πρόσωπα απομονώθηκαν από τις υπόλοιπες Wikipedia οντότητες, δημιουργήθηκαν στη ΒΔ προσωρινοί πίνακες, ένας για καθεμιά από τις προς μελέτη Wikipedia versions, με όνομα της μορφής: TEMP_3XX0 (TEMP_v3510, TEMP_v3600, TEMP_v3800)
και στήλες (όνομα οντότητας, κατηγορία οντότητας,όνομα ιδιότητας, ακατέργαστη τιμή ιδιότητας).
Καθένας από τους προσωρινούς αυτούς πίνακες γέμισε, με νέο διάβασμα των αποσυμπιεσμένων Raw Infobox Properties αρχείων, με τα στιγμιότυπα (βλ. Ορισμός 4 κεφ. 3.1 και κεφ. 4.2.1) των προσώπων που απαντούν στην αντίστοιχη Wikipedia έκδοση αλλά και στον πίνακα TEMP_persons568, που προηγουμένως δημιουργήθηκε.
Σε αυτή τη ΒΔ θα αναφερόμαστε στο εξής με το χαρακτηριστικό $\mathbf{ΒΔ_t}$, όπου ο δείκτης t υποδηλώνει την κατάσταση της βάσης τη χρονική στιγμή που έχει περατωθεί η διαδικασία της εξόρυξης των προσώπων και έχουν γεμίσει οι προσωρινοί πίνακες κάθε έκδοσης.
Η υλοποίηση αυτή (κώδικας Python v.3.3 versionLoader.py) αποτελεί επίσης τμήμα του component Persons' Data miner, που περιγράφεται στο κεφ. 5.1

4.2.4 Δημιουργία χρονικά ευαίσθητης ΒΔ με Οντότητες - Entities' DataBase

Στη συνέχεια με ένα σύστημα sqls commands επί της $\mathbf{ΒΔ_t}$, που δημιουργήθηκε στο βήμα 4.2.3, δημιουργήθηκαν:

- **ένας πίνακας persons** με στήλες: pid (unique), name (unique) **που περιέχει όλες εν γένει τις οντότητες-πρόσωπα** (κατά τον Ορισμό 1 του κεφ. 3.1 του παρόντος) που εμφανίζονται στο σύστημα της Wikipedia ανεξαρτήτως έκδοσης
- **ένας πίνακας category** με στήλες: cid (unique), name (unique) **που περιέχει όλες εν γένει τις κατηγορίες οντοτήτων** (βλ. κεφ. 3.1 του παρόντος) που εμφανίζονται στο σύστημα της Wikipedia ανεξαρτήτως έκδοσης.
- **ένας πίνακας percat** με στήλες: percatid, pid (με αναφορά στον πίνακα persons), cid (με αναφορά στον πίνακα category) **που αποδίδει την ένταξη της οντότητας σε κατηγορία.**
- **ένας πίνακας properties** με στήλες: propid (unique), propname, **που περιέχει όλες εν γένει τις ιδιότητες των οντοτήτων**(κατά τον Ορισμό 2 του κεφ. 3.1 του παρόντος)

που εμφανίζονται στο σύστημα της Wikipedia ανεξαρτήτως έκδοσης, κατά τα οριζόμενα στο κεφ. 3.1 του παρόντος.

- **ένας πίνακας perprop** με στήλες: perpropid(unique), pid (αναφορά στον πίνακα persons), propid (αναφορά στον πίνακα properties) **που περιέχει μέγιστο δυνατό σύνολο P των ιδιοτήτων κάθε συγκεκριμένης οντότητας-προσώπου** στις υπό μελέτη εκδόσεις**,** κατά τα οριζόμενα στο κεφ. 3.1 του παρόντος.
- **ένας πίνακας (versionPrefix)_howmany** με στήλες: perpropid (unique) (αναφορά στον πίνακα perprop) και howmany, **που περιέχει την πληθυκότητα** (κατά τα περιγραφόμενα στο κεφ. 3.2 του παρόντος) **των χαρακτηριστικών μιας οντότητας.**
- ένας **πίνακας (versionPrefix)_perpropval** με στήλες: perpropid (unique) (αναφορά στον πίνακα perprop), rawval, fineval **που περιέχει τα χαρακτηριστικά των οντοτήτων και άρα τα στιγμιότυπά τους** (κατά τον Ορισμό 5 του κεφ. 3.1 του παρόντος)

  Στη στήλη rawval του πίνακα αυτού γράφτηκε η ακατέργαστη τιμή της ιδιότητας, όπως ακριβώς εμφανίζεται στο αρχείο RawInfoboxProperties, ενώ στην στήλη fineval η τιμή που εξορύχθηκε από την ακατέργαστη αυτή τιμή (όπως περιγράφεται αναλυτικά στο Παράρτημα Β2)

Μετασχηματίσαμε έτσι τη $ΒΔ_t$ που περιγράφεται στο 4.2.3 σε μια **χρονικά "ευαίσθητη" ΒΔ οντοτήτων** (**Temporal Entities' DataBase - TEDB**), στην οποία υπάρχουν: οντότητες (table:persons), κατηγορίες (table:category), membership relationships (table:percat), ιδιότητες (table:properties), χαρακτηριστικά με την πληθυκότητα τους για κάθε συγκεκριμένη έκδοση (table: vXXXX_howmany), σύνολα P ιδιοτήτων οντότητας (table:perprop), στιγμιότυπα των προσώπων σε συγκεκριμένη χρονική περίοδο αναφοράς, που αντιστοιχεί στην κάθε έκδοση (table:vXXXX_perpropval) κατά τα οριζόμενα στο 3.1 του παρόντος.

Η υλοποίηση της TEDB αποτελεί τμήμα του component Entities Creator που Περιγράφεται στο κεφ. 5.1 *(κώδικας Python v.3.3 versionLoader.py)*

Πρέπει να διευκρινιστεί ότι, για καθαρά πρακτικούς λόγους, ΔΕΝ χρησιμοποιήσαμε δύο διαφορετικές βάσεις σε ΦΥΣΙΚΟ ΕΠΙΠΕΔΟ. Η βάση στην οποία δουλέψαμε ήταν πάντα η SQLite3 βάση δεδομένων C:\webapp\data\ dbpediaPersons.sqlite που ξεκίνησε από τη δομή $ΒΔ_t$ και, εμπλουτισμένη, κατέληξε τη χρονική στιγμή t+Δt, κατά την οποία είχε περατωθεί η επεξεργασία της $ΒΔ_t$ κατά τα αναφερόμενα στο 4.2.4, στη δομή TEDB, που ισοδυναμεί με την $ΒΔ_{t+Δt}$. Βεβαίως θα μπορούσαμε να έχουμε διαγράψει από το σχήμα της (Data Base schema) όλους τους προσωρινούς βοηθητικούς πίνακες. Αυτό όμως δε θα ήταν καθόλου πρακτικό, καθώς τα Raw

Infobox Properties files είναι τεράστια (50.000.000 lines το κάθε ένα κατά μέσο όρο) και θα ήταν χρονοβόρα η επανεπεξεργασία τους κατά την ανάπτυξη του συστήματος (testing, debugging).
Τα sql commands τα οποία δημιούργησαν και γέμισαν τους πίνακες της προσωρινής $ΒΔ_t$, καθώς και της TEDB από τους αντίστοιχους temporary για κάθε version (v3510_TEMP, v3600_TEMP v3800_TEMP), εμφανίζονται στο ΠΑΡΤΗΜΑ Β1.

4.3 Επεξεργασίες κατά τη δημιουργία της TEDB

4.3.1 Το πρόβλημα με τις πολλαπλές αναφορές κατηγοριών.

Πριν την εισαγωγή τους στον αντίστοιχο πίνακα (category) και μετά από προσεκτική παρατήρηση, διαπιστώθηκε ότι τα ονόματα των κατηγοριών των προσώπων δεν δηλώνονταν πάντα με τον ίδιο τρόπο. Π.χ. η κατηγορία actor μπορεί να εμφανιζόταν ως Actor αλλά και ως actor, ενώ π.χ. η κατηγορία Hockey_Player εμφανιζόταν και ως hockey_player, Hockey_player, hockey_Player. Για την αντιμετώπιση του προβλήματος αυτού αποφασίστηκε να τηρηθεί η σύμβαση των πεζών σε όλους τους χαρακτήρες του λεκτικού της κατηγορίας. Έτσι, στον πίνακα category της TEDB, εισήχθηκαν οι διακριτές τιμές ονομάτων κατηγοριών που προέκυψαν, αφού πρώτα μετατράπηκαν σε πεζούς όλοι οι χαρακτήρες των ονομάτων των κατηγοριών. Με αυτόν τον τρόπο αποφεύχθηκαν πολλαπλές αναφορές.

4.3.2 Το πρόβλημα με τις πολλαπλές αναφορές ιδιοτήτων.

Κατά την επεξεργασία των δεδομένων διαπιστώθηκε ότι πολλές διαφορετικές ιδιότητες μοντελοποιούσαν την ίδια έννοια του φυσικού κόσμου.
Για παράδειγμα: στο πρόσωπο “Aristotle Onassis” στη version 3.6 υπάρχει η ιδιότητα “dateofdeath” αλλά και η “deathdate”, και οι δύο αναφορές στην ημερομηνία θανάτου του. Ομοίως για το πρόσωπο “Abraham Lincoln” στην version 3.6 υπάρχει η ιδιότητα “placeofbirth” και “birthplace”, και οι δύο αναφορές στον τόπο γέννησής του.
Η αντιμετώπιση (matching) του προβλήματος αυτού, η οποία δεν είναι τόσο απλή, θεωρήθηκε ότι δεν εμπίπτει στα όρια της μεταπτυχιακής διατριβής αυτής.

4.4 Ερευνητικά ερωτήματα

4.4.1 Κριτήριο Ομαδοποίησης.

Ένας από τους τελικούς στόχους της παρούσας μεταπτυχιακής διατριβής, όπως ορίστηκαν στο κεφ. 1.2 του παρόντος, ήταν η δημιουργία μιας χρονικά ευαίσθητης Βάσης Δεδομένων, η οποία θα μπορεί να απαντά σε ερωτήματα (queries), των οποίων τα αποτελέσματα θα οδηγούν στην εξαγωγή συμπερασμάτων για την εξέλιξη των οντοτήτων ατομικά ή/και συλλογικά. Γι' αυτό ιδιαίτερη προσοχή δόθηκε στο σχεδιασμό των ερωτημάτων αυτών.

Τα ερευνητικά ερωτήματα (queries) ομαδοποιήθηκαν ανάλογα με το πεδίο εφαρμογής τους σε:

- **Γενικά**, που εξάγουν γενικές παρατηρήσεις για τις οντότητες ή ομάδες οντοτήτων μιας Wikipedia έκδοσης (version) (τύπου **G**eneral)
- **Συγκριτικά**, που συγκρίνουν γενικές τάσεις ανάμεσα στους κόσμους δύο (και όχι περισσότερων) Wikipedia εκδόσεων (τύπου **C**omparative).
- **Ερωτήματα για συγκεκριμένο πρόσωπο** που :
    - είτε το μελετούν σε διάφορες χρονικές περιόδους, δηλ. σε δύο διαφορετικές versions (τύπου **P**erson**C**omparative, **PC**)
    - είτε μελετούν τις ιδιότητές του σε μια version (τύπου **P**erson**G**eneral, **PG**)

Η ομαδοποίηση αυτή κρίθηκε σκόπιμη, διότι έτσι εξασφαλίζεται η θεώρηση της εξέλιξης από διαφορετικές οπτικές γωνίες:

- της ατομικής ενός συγκεκριμένου προσώπου (π.χ ποια χαρακτηριστικά είχε ο Αριστοτέλης Ωνάσης στην έκδοση 3.8),
- της ομαδικής για κατηγορίες προσώπων (π.χ. πόσες ιδιότητες αποδίδονται στους μουσικούς στην έκδοση 3.6 και πόσες στην 3.8) και
- της συνολικής που αφορά στον κόσμο μιας ολόκληρης Wikipedia έκδοσης (π.χ. ποια πρόσωπα που υπάρχουν στην Wikipedia 3.8 δεν υπήρχαν στην 3.6)

Δεν μπορούσαμε επίσης να αφαιρέσουμε τη δυνατότητα από το χρήστη του συστήματος (πρόσωπο ή άλλο σύστημα) να λάβει πληροφορίες για πρόσωπα, ιδιότητες, κατηγορίες και χαρακτηριστικά μιας συγκεκριμένης έκδοσης της Wikipedia, αν αυτό επιθυμεί (τύπος G).

Η υλοποίηση της αποστολής ερευνητικών ερωτημάτων στην TEDB (κώδικας Python v.3.3loadQueries.py) και η λήψη αποτελεσμάτων (κώδικας Python v.3.3 getResults.py) εντάσσονται στο component WebApp που περιγράφεται στο κεφ. 5.1.

4.4.2 Παράθεση των ερευνητικών ερωτημάτων.

Στις παρακάτω παραγράφους παραθέτουμε τα ερευνητικά ερωτήματα (queries) που επιλέχτηκαν να χρησιμοποιηθούν σε αυτήν τη μεταπτυχιακή διατριβή, ομαδοποιημένα ανάλογα με την κατηγορία στην οποία ανήκουν.

Τα ερωτήματα τύπου G*(eneral)*:

G 1. Πόσα αντικείμενα (Πρόσωπα) έχει η επιλεγμένη Wikipedia version

G 2. Πόσες διαφορετικές ιδιότητες εμφανίζονται στην επιλεγμένη Wikipedia version

G 3. Πόσες διαφορετικές κατηγορίες εμφανίζονται στην επιλεγμένη Wikipedia version

G 4. Κατηγορίες της επιλεγμένης Wikipedia version κατά φθίνουσα σειρά πλήθους ιδιοτήτων.

G 5. Πρόσωπα στην επιλεγμένη Wikipedia version κατά φθίνουσα σειρά πλήθους ιδιοτήτων.

G 6. Ποιες είναι οι πολυπληθέστερες κατηγορίες στην επιλεγμένη Wikipedia version

G 7. Ποια πρόσωπα από κάθε κατηγορία έχουν τις περισσότερες ιδιότητες; (δημοφιλέστερα;)

G 8. Ποιες ιδιότητες υπάρχουν στην επιλεγμένη Wikipedia version

Τα ερωτήματα τύπου C*(omparative)*:

C1. Πώς εξελίσσεται το πλήθος των οντοτήτων (προσώπων) στις επιλεγμένες Wikipedia versions, δηλαδή πόσες οντότητες είχαμε σε μια version και πόσες στην άλλη

C2. Πώς εξελίσσεται το πλήθος των ιδιοτήτων στις επιλεγμένες Wikipedia versions, δηλαδή πόσες ιδιότητες είχαμε σε μια version και πόσες στην άλλη

C3. Ποιες οντότητες (πρόσωπα) προστέθηκαν από την επιλεγμένη παλαιότερη Wikipedia version προς την επιλεγμένη νεότερη της

C4. Ποιες οντότητες (πρόσωπα) διαγράφηκαν από την επιλεγμένη παλαιότερη Wikipedia version προς την επιλεγμένη νεότερή της

C5. Ποιες ιδιότητες, που υπάρχουν στην επιλεγμένη νεότερη Wikipedia version εμφανίζονται για πρώτη φορά

C6. Ποιες ιδιότητες,που υπάρχουν στην επιλεγμένη νεότερη Wikipedia version, δεν υπήρχαν στην επιλεγμένη προηγούμενη

C7. Κατά πόσο αυξήθηκαν/μειώθηκαν οι ιδιότητες των οντοτήτων (Προσώπων) μεταξύ των 2 επιλεγμένων Wikipedia versions

C 8. Πώς εξελίσσεται το πλήθος των κατηγοριών στις επιλεγμένες Wikipedia versions

C 9. Ποιες κατηγορίες υπάρχουν στην παλαιότερη και ποιες στην νεότερη Wikipedia version

C10. Πώς εξελίσσεται το συνολικό πλήθος των ιδιοτήτων των προσώπων ανά κατηγορία μεταξύ των 2 επιλεγμένων Wikipedia versions

C11. Πώς εξελίσσεται το συνολικό πλήθος των προσώπων ανά κατηγορία μεταξύ των 2 επιλεγμένων Wikipedia versions

C12. Σε ποιες οντότητες (πρόσωπα) προστέθηκαν πάνω από μία τιμές (και πόσες) από τη μια στην άλλη επιλεγμένη Wikipedia version

C13. Σε ποιες οντότητες (πρόσωπα) αφαιρέθηκαν πάνω από μία τιμές (και πόσες) από τη μια στην άλλη επιλεγμένη Wikipedia version

C14. Ποιων προσώπων (filter:pid modulo 11) οι ιδιότητες - κοινές στις 2 επιλεγμένες Wikipedia versions και μονότιμες σε καθεμιά - έχουν διαφορετική τιμή στη νεότερη version και ποιες είναι αυτές

C15. Ποιο Πρόσωπο έχει τις περισσότερες ιδιότητες σε όλες τις Wikipedia versions και σε ποια κατηγορία ανήκει

C16. Πώς μεταβάλλονται τα πρόσωπα με τις περισσότερες ιδιότητες (δημοφιλέστερα;), ανά κατηγορία από την επιλεγμένη παλαιότερη Wikipedia version στη νεότερη;

Τα ερωτήματα τύπου P(*erson*)C(*omparative*):

PC1. Ποιες ιδιότητες έχει ο/η (πρόσωπο επιλογής) σε καθεμιά από τις επιλεγμένες Wikipedia versions

PC2. Ποιες ιδιότητες του/της (πρόσωπο επιλογής), μονότιμες και κοινές στις επιλεγμένες Wikipedia versions, έχουν διαφορετική τιμή στη νεότερη version

PC3. Ποιες είναι οι τιμές των ιδιοτήτων του/της (πρόσωπο επιλογής) που είναι μονότιμες και στις 2 επιλεγμένες Wikipedia versions, καθεμιά από αυτές

PC4. Πόσες ιδιότητες έχει ο/η (πρόσωπο επιλογής) στην παλαιότερη Wikipedia version και πόσες στη νεότερη

PC5. Ποιες ιδιότητες του/της (πρόσωπο επιλογής) είναι καινούργιες στην νεότερη Wikipedia version

PC6. Ποιες ιδιότητες του/της (πρόσωπο επιλογής) διαγράφηκαν από τη νεότερη Wikipedia version, ενώ υπήρχαν στην παλαιότερη

PC7. Ποια είναι τα χαρακτηριστικά του/της (πρόσωπο επιλογής) ανύπαρκτα στην επιλεγμένη παλαιότερη Wikipedia version, που προστέθηκαν στην επιλεγμένη νεότερη

PC8. Ποια χαρακτηριστικά του/της (πρόσωπο επιλογής), υπαρκτά στην επιλεγμένη παλαιότερη Wikipedia version, διαγράφηκαν από την επιλεγμένη νεότερη

Τα ερωτήματα τύπου P(*erson*)G(*eneral*):

PG1. Ποιο είναι το πλήθος ιδιοτήτων του προσώπου (πρόσωπο επιλογής) στην Wikipedia version (έκδοση επιλογής)

PG2. Ποια είναι τα χαρακτηριστικά του/της (πρόσωπο επιλογής) στην Wikipedia version (έκδοση επιλογής)

4.4.3 Σχολιασμός / Χρησιμότητα των ερευνητικών ερωτημάτων

Τα ερωτήματα που επεξεργαζόμαστε, και πιο συγκεκριμένα η παρακολούθηση ενός συστήματος για την ανάκτηση των αποτελεσμάτων των ερωτημάτων, μπορούν να χρησιμοποιηθούν από αλγορίθμους, που χειρίζονται τέτοιου είδους ευμετάβλητα δεδομένα για την βελτίωση της λειτουργίας τους (όπως έχουμε περιγράψει και στο κεφ.1.1) αλλά και από άλλους χρήστες (συστήματα ή πρόσωπα) για περαιτέρω μελετητική δουλειά. Στις παρακάτω παραγράφους προτείνουμε πιο συγκεκριμένα πεδία χρησιμότητας μερικών από τα ερωτήματα αυτά.

Το ερώτημα G3 σε συνδυασμό με το ερώτημα G6 θα μπορούσε να μας οδηγήσει στο να χρησιμοποιούμε διαφορετικούς πόρους για αποθήκευση και διαχείριση των πολυπληθών κατηγοριών. Για παράδειγμα, αν διαπιστώνεται ότι η κατηγορία «μουσικοί» αριθμεί τα περισσότερα μέλη, τότε θα μπορούσαμε να αποθηκεύσουμε τα στιγμιότυπα των μουσικών σε χωριστή ΒΔ και να έχουμε έναν εξυπηρετητή αποκλειστικά γι αυτή την κατηγορία. Επίσης, καθώς και η ίδια η κατηγορία δεν αποτελεί παρά μια ιδιότητα (του μέλους σε αυτή την κατηγορία) της οποίας η τιμή είναι η ίδια για πολλά πρόσωπα, η μελέτη των κατηγοριών μπορεί να βοηθήσει στον εντοπισμό σχέσεων μεταξύ των οντοτήτων.

Το G4 μπορεί να μας δώσει ποιοτική πληροφορία, η οποία να αξιοποιηθεί για περαιτέρω μελέτη των οντοτήτων – μελών μιας συγκεκριμένης κατηγορίας είτε μόνο του είτε (και καλύτερα) σε συνδυασμό με το G6. Για παράδειγμα, αν διαπιστώνεται ότι μια κατηγορία έχει πάρα πολλές ιδιότητες αλλά πολύ λίγα μέλη, μπορεί να θέλουμε να τη μεταχειριστούμε διαφορετικά ή να αναζητήσουμε τους λόγους για τους οποίους έχει τόσες πολλές ιδιότητες (ενδεχόμενες πολλαπλές αναφορές/ multiple references, φαινόμενο στο οποίο αναφερθήκαμε εν συντομία στο κεφ. 4.3.2) Αν πάλι μια κατηγορία είναι πολυπληθέστατη και ταυτόχρονα τα μέλη της περιγράφονται από πολλές ιδιότητες, τότε μπορεί να αξίζει να σκεφτούμε τρόπους προτυποποίησης της απόδοσης ιδιοτήτων (ώστε να αποφύγουμε πολλαπλές αναφορές) ή/και συγχώνευσης ιδιοτήτων.

Ερωτήματα της μορφής των G5 και C15 αλλά και G7 θα μπορούσαν να χρησιμεύσουν στο να έχουμε σε πρώτη ζήτηση πρόσωπα που αλλάζουν συχνά (ώστε π.χ. να μην χρειάζονται για αυτά πολλά I/O στην TEDB). Κι αυτό με το σκεπτικό ότι, εάν ένα πρόσωπο απέκτησε πολλές ιδιότητες, (ενδέχεται να) σημαίνει ότι οι χρήστες το αναζητούν συχνά, με σκοπό να το τροποποιήσουν. Ενδέχεται, διότι θα μπορούσαν όλες αυτές οι ιδιότητες να προστέθηκαν μεμιάς από ένα χρήστη κατά την διάρκεια μιας session. Επομένως τα αποτελέσματα των παραπάνω ερωτημάτων μόνο ενδεικτικά μπορεί να λειτουργήσουν και για την πλήρη αξιοποίησή τους χρειάζονται και άλλου είδους δεδομένα (π.χ. από log files).

Ερωτήματα του τύπου C1,C2,C8, C12, C13 και C14 οδηγούν σε εκτιμήσεις του ρυθμού μεταβολής, ο οποίος με τη σειρά του μπορεί να οδηγήσει σε εκτιμήσεις για αποθηκευτικούς χώρους και συνεπαγόμενα κόστη. Ειδικά μάλιστα το ερώτημα C14 μπορεί να μας οδηγήσει να διαχειριστούμε με διαφορετικό τρόπο ιδιότητες των προσώπων οι οποίες αλλάζουν πιο συχνά από άλλες που δεν αλλάζουν. Για παράδειγμα, οι πληροφορίες για την ημερομηνία γέννησης ενός ατόμου αλλάζουν αραιότερα από άλλες που αφορούν - ας πούμε - την οικογενειακή του κατάσταση. Έτσι για τον Brad Pitt έχουμε αμετάβλητη ημερομηνία γέννησης, αλλά αλλαγές στις ιδιότητες spouse, partner, domesticpartner.

Το αποτέλεσμα του ερωτήματος C10 (σε συνδυασμό με τα αποτελέσματα του C11) μπορεί να μας οδηγήσει σε αποφάσεις για διαχωρισμό ορισμένων κατηγοριών . Κατηγορίες που τα μέλη τους και το πλήθος των ιδιοτήτων τους αλλάζουν με γοργούς ρυθμούς θα πρέπει να κατευθυνθούν προς διαφορετικό είδος διαχείρισης, έτσι ώστε η συνολική εμπειρία του χρήστη να βελτιωθεί. Για παράδειγμα, στην κατηγορία μουσικοί καλλιτέχνες προστέθηκαν 4331 πρόσωπα από την version v 3.5.1 στη χρονικά επόμενή της, την 3.6, και από αυτήν μέχρι την μεθ-επόμενή της και τρέχουσα (3.8) προστέθηκαν άλλα 8884 άτομα, ενώ οι κάτοχοι ψηλών θέσεων στην κρατική διοίκηση (office holders) σημειώνουν διαφορά στις ίδιες versions 13906 ατόμων!! Αντίστοιχα 153 επιπλέον ιδιότητες εισήλθαν στην κατηγορία μουσικοί (προσοχή:ιδιότητες όχι χαρακτηριστικά, δηλ. ζεύγος ιδιότητας - τιμής) από την 3.5.1 στην 3.6 και 287 επιπλέον εισήλθαν στην ίδια κατηγορία από την 3.6 στην 3.8.

Τα ερωτήματα C3 και C4 μπορεί να ενδιαφέρουν εφαρμογές που συσχετίζουν στιγμιότυπα (lineage) ή που παρακολουθούν την πορεία των οντοτήτων στο χρόνο. Μια οντότητα μπορεί να διαγράφηκε, γιατί αντικαταστάθηκε από ή εξελίχθηκε σε μια άλλη και μια οντότητα μπορεί να προστέθηκε, γιατί προέκυψε (εξελίχθηκε από κάποια προηγούμενη).

Αντίστοιχα ερωτήματα του τύπου C5, C6 μπορούν να βοηθήσουν εφαρμογές οι οποίες μελετούν σχέσεις μεταξύ εννοιών μια και κάθε ιδιότητα υποκρύπτει μια σχέση ανάμεσα σε δύο έννοιες[1]. Επίσης μπορούν να δώσουν προβλέψεις, καθώς, όταν μια ιδιότητα εμφανίζεται για πρώτη φορά σε μια version ενώ πριν δεν υπήρχε, είναι πιθανόν να διαχυθεί αργότερα στο σύνολο της οντολογίας. (παράδειγμα: skypeAccount)

Τέλος, όλα τα ερωτήματα που αφορούν συγκεκριμένο πρόσωπο (τύπος PC, PG) μπορούν να αξιοποιηθούν από εφαρμογές που μελετούν και ενσωματώνουν δεδομένα από διαφορετικές πηγές και συσχετίζουν στιγμιότυπα οντοτήτων μεταξύ τους. Τα αποτελέσματα ερωτημάτων τύπου PG μπορεί να χρησιμεύσουν στην βελτίωση συστημάτων που διαχειρίζονται οντότητες. Για παράδειγμα ένα τέτοιο σύστημα θα μπορούσε να ενδιαφέρεται για τα χαρακτηριστικά μιας

οντότητας μια συγκεκριμένη χρονική στιγμή (που αντιστοιχεί σε μια συγκεκριμένη Wikipedia version).

Ειδική μνεία θα κάνουμε:

- στο G8 στου οποίου τα αποτελέσματα έχουμε την δυνατότητα, με τοπική αναζήτηση στην σελίδα, να διαπιστώσουμε εύκολα φαινόμενα πολλαπλών αναφορών. Για παράδειγμα εισάγοντας τον όρο death βλέπουμε ότι μπορούμε να βρούμε τα 'ageatdeath', 'ageofdeath' που αμφότερα αναφέρονται στην ηλικία θανάτου, αλλά και 'countryofdeath','countryofdath', 'countryofdaeth' που αναφέρονται στην χώρα θανάτου και 'datedeath', 'dateofdeath', dateofdeathdate' που αναφέρονται στην ημερομηνία θανάτου.
- Στο C9 το οποίο μας επιτρέπει παρατηρήσεις επί της εξέλιξης της οντολογίας της Wikipedia μεταξύ των versions. (βλ. και κεφ. Μεθοδολογία.)

4.4.4 Η Υλοποίηση με τον πίνακα queries.

Τα ερευνητικά ερωτήματα (queries) υλοποιήθηκαν ως sql statements. Καθώς όμως ένας από τους κεντρικούς στόχους της παρούσας μεταπτυχιακής διατριβής ήταν η δημιουργία ενός συστήματος (web application) το οποίο θα έδινε την δυνατότητα **στο χρήστη** (πρόσωπο ή άλλο σύστημα) να παρακολουθεί την εξέλιξη οντοτήτων και ομάδων οντοτήτων μεταξύ δύο Wikipedia versions, αλλά και τα χαρακτηριστικά των ίδιων των εκδόσεων ως δυνατών κόσμων, με απλό και κατανοητό τρόπο, κατά την άποψή μας έπρεπε:

- Να υπάρχει η δυνατότητα δυναμικής αλλαγής των ερευνητικών αυτών ερωτημάτων π.χ. κατάργηση ενός ερωτήματος ή εισαγωγή ενός άλλου, που όμως να μην οδηγεί σε αλλαγή του κώδικα υλοποίησης
- Να μπορεί η εμφάνιση των αποτελεσμάτων στο χρήστη να προσαρμόζεται δυναμικά στα ιδιαίτερα χαρακτηριστικά κάθε ερωτήματος (π.χ. πλήθος στηλών, επικεφαλίδες)
- Να παρέχονται πληροφορίες στο χρήστη του (πρόσωπο ή άλλο σύστημα) για τη συνεισφορά τού κάθε ερευνητικού ερωτήματος στην επίτευξη του δεύτερου από τους κεντρικούς στόχους της μεταπτυχιακής διατριβής αυτής (βλ. κεφ. 1.2), που ήταν η παρακολούθηση της εξέλιξης των αντικειμένων στο Διαδίκτυο

Έτσι κάθε ερευνητικό ερώτημα καταχωρίστηκε σε ένα πίνακα queries, που δημιουργήθηκε στην TEDB, με στήλες:

1. qcat (κατηγορία ερευνητικού ερωτήματος, G,C,P) primary-key
2. qid (autoincrement) primary-key
3. qsql (sql statement)

4. qdesc_el (τίτλος στα Ελληνικά του ερευνητικού ερωτήματος)
5. qdesc_en (τίτλο στα Αγγλικά του ερευνητικού ερωτήματος)
6. outputTableHeaders_el (semicolon delimited list, με επικεφαλίδες στα Ελληνικά των στηλών που επιστρέφει το sql statement ως output)
7. outputTableHeaders_en (semicolon delimited list, με επικεφαλίδες στα Αγγλικά των στηλών που επιστρέφει το sql statement ως output)
8. quse_el (κείμενο στα Ελληνικά για τη συνεισφορά του συγκεκριμένου ερωτήματος στην επίτευξη του στόχου της μεταπτυχιακής διατριβής)
9. quse_en (κείμενο στα Αγγλικά για τη συνεισφορά του συγκεκριμένου ερωτήματος στην επίτευξη του στόχου της μεταπτυχιακής διατριβής)

Η υλοποίηση αυτή παρέχει την δυνατότητα της άμεσης εισόδου στην λειτουργικότητα του συστήματος οποιουδήποτε ερευνητικού ερωτήματος κριθεί ότι πρέπει να προστεθεί, διότι προάγει την έρευνα, με μια κατάλληλη καταχώρισή του στον πίνακα queries

Το sql statement που δημιούργησε τον πίνακα queries εμφανίζεται στον αριθμό 24 του παραρτήματος Β1. Για την εισαγωγή ενός νέου ερευνητικού ερωτήματος στο σύστημα αναλυτικές οδηγίες παρέχονται στο Παράρτημα Γ.

4.5 Ενσωμάτωση των επιλογών του χρήστη στα SQL statements και στο User Interface

Στην στήλη qsql του πίνακα queries περιέχεται, όπως περιγράφηκε πιο πάνω, το sql statement, που υλοποιεί κάθε ερευνητικό ερώτημα, σε text μορφή. Σε αυτό το sql statement υπεισέρχονται πίνακες. Όμως στην TEDB υπάρχουν πίνακες, όπως π.χ οι persons, properties, category, στους οποίους θα αναφερθούμε, κατά την σύνταξη ενός sql statement, ανεξάρτητα από την έκδοση ή τις εκδόσεις που μελετάμε και πίνακες, όπως π.χ ο v3510_howmany, που αφορούν μια συγκεκριμένη έκδοση(version) και στους οποίους θα αναφερθούμε μόνο αν μελετάμε την συγκεκριμένη έκδοση.
Το σύστημα προφανώς δεν μπορεί εκ των προτέρων να γνωρίζει ποια ή ποιες έκδοση/εκδόσεις θα επιλέξει ο χρήστης να μελετήσει, μπορεί όμως να προσαρμοστεί δυναμικά όταν το μάθει.
Έτσι κάθε πίνακας εξαρτώμενος από Wikipedia έκδοση εμφανίζεται στο sql statement με κάποια από τις 2 μορφές που ακολουθούν, ανάλογα με την "χρονική θέση" του στο σύστημα.

- είτε @_howmnay (αντί π.χ. v3510_howmany)

- είτε #_howmnay (αντί για π.χ. v3800_howmany)

Συγκεκριμένα η σύμβαση που τηρήθηκε κατά τη δημιουργία των SQL statements είναι η ακόλουθη:

- όταν ΕΝΑΣ (αριθμός 1) πίνακας υπεισέρχεται στο sql statement (σε ερωτήματα τύπου G, PG), τότε η επιλογή του χρήστη, όπως αντλείται από το πρόγραμμα, αντικαθιστά το @
- όταν ΔΥΟ[1] πίνακες υπεισέρχονται στο sql statement (σε ερωτήματα τύπου C , PC), τότε οι επιλογές του χρήστη, όπως αντλούνται από το πρόγραμμα, αντικαθιστούν τα @ και #. Η νεότερη version αντικαθιστά το # και η παλαιότερη το @.

Επίσης, στο κείμενο (text) κάθε SQL statement, στο οποίο υπεισέρχεται το όνομα κάποιου προσώπου ως παράμετρος (προκειμένου για ερωτήματα που αφορούν σε συγκεκριμένο πρόσωπο), που είναι καταχωρισμένο στην στήλη qsql του πίνακα queries, χρησιμοποιείται placeholder στη θέση του ονόματος του προσώπου ως εξής :

" ……..where pname like"%?%"……".

Όταν ο χρήστης επιλέξει το υπό μελέτη πρόσωπο, το placeholder "?" αντικαθίσταται, on the fly, από την επιλογή του.

Τα παραπάνω φαίνονται στην εικόνα 4.1, που ακολουθεί.

```
select prop.propname Property,a1.fineval OldValue,a2.fineval newValue
from
 (select a.perpropid from @_howmany a join #_howmany b on a.perpropid = b.perpropid where a.howmany = 1 and b.howmany = 1) w
join @_perpropval a1 on    w.perpropid = a1.perpropid
join  #_perpropval a2 on a1.perpropid = a2.perpropid
join perprop pp on pp.perpropid= a1.perpropid
join persons p on  p.pid = pp.pid
join properties prop on prop.propid = pp.propid
WHERE p.pname = "?" and    lower(replace(a1.fineval,'_',' ')) <> lower(replace(a2.fineval,'_',' ') )
```

Εικ. 4.1 Sql statement του πίνακα queries, στο οποίο φαίνονται όλες οι συμβάσεις της υλοποίησης.

[1] Σημειώνουμε ότι οι συγκρίσεις γίνονται πάντα ανά δύο Wikipedia εκδόσεις δηλ. σε δύο χρονικές περιόδους.

Κεφάλαιο 5
Το σύστημα WOE

Στο κεφάλαιο αυτό θα παρουσιαστεί το σύστημα **WOE**, από τα αρχικά των λέξεων **W**ikipedia **W**eb **O**bjects' **E**volution, που αναπτύχθηκε στα πλαίσια της μεταπτυχιακής διατριβής αυτής. Πρόκειται για ένα εργαλείο με το οποίο ο ενδιαφερόμενος χρήστης μπορεί, χρησιμοποιώντας μια TEDB (βλ.κεφ. 4.2.4 και 4.3), που το ίδιο το σύστημα WOE δημιουργεί βασισμένο σε επιλογές του, εύκολα να μελετήσει τις αλλαγές που επισυνέβησαν στο χρόνο που μεσολάβησε ανάμεσα σε δύο διαφορετικές Wikipedia εκδόσεις σε :

- πρόσωπα,
- ομάδες προσώπων
- τις ίδιες τις εκδόσεις ως δυνατούς κόσμους.

Πρέπει να σημειωθεί ότι η παρουσίαση της ίδιας οντότητας ή/και ομάδας οντοτήτων (κατηγορίας) σε διαφορετικές χρονικές στιγμές συνιστά καινοτομία στο πεδίο.

Στο κεφ. 5.1 παρουσιάζεται η αρχιτεκτονική του συστήματος και τα εργαλεία με τα οποία υλοποιήθηκε, στο κεφ. 5.2 οι προσφερόμενες από αυτό λειτουργίες, στο κεφ. 5.3 οι μελλοντικές επεκτάσεις του, ενώ στο κεφ. 5.4 γίνεται μια σύνοψη συμπερασμάτων.

5.1 Αρχιτεκτονική του συστήματος

Όλα τα components του συστήματος υλοποιήθηκαν με λογισμικό σε γλώσσα προγραμματισμού Python v.3.3, που στη default εγκατάστασή της ενσωματώνει SQLite3 Βάση Δεδομένων, η οποία και χρησιμοποιήθηκε, και server που αποτέλεσε τον εξυπηρετητή για τη web application που αναπτύξαμε.

Το σύστημα WOE αποτελείται από τρία (3) κύρια μέρη (components).

1. **Το component PDM** (**Person's Data Miner**), που χειρίζεται την εξόρυξη των προσώπων, αλλά και των στιγμιοτύπων τους από τα Raw Infobox Properties files των υπό μελέτη Wikipedia versions (βλ. κεφ. 4.2.2).

Το component αυτό υλοποιεί το download, την αποσυμπίεση και την πρώτη επεξεργασία των αρχείων που περιέχουν τα σύνολα δεδομένων των Wikipedia εκδόσεων (Raw Infobox Properties.bz2) με σκοπό:

- τον εντοπισμό και την απομόνωση των προσώπων, καθώς και
- τη δημιουργία μιας σχεσιακής βάσης με προσωρινούς πίνακες με τα δεδομένα των προσώπων αυτών σε κάθε προς μελέτη Wikipedia έκδοση.

Για αυτό και εκτελεί την εργασία του σε 2 χρόνους:

Πρώτα εντοπίζει τα πρόσωπα και τα γράφει στο σχετικό πίνακα (TEMP_persons568, βλ. κεφ. 4.2.3). Η γνώση αυτή τού είναι απαραίτητη, ώστε να μπορέσει, στη συνέχεια, να αποσπάσει από τα αρχεία που διαβάζει μόνο τις γραμμές εκείνες που αφορούν πρόσωπα..

Για να επιτύχει τον εντοπισμό αυτόν, είναι εφοδιασμένο με μια σειρά από κανόνες, οι οποίοι προκύπτουν από την αξιοποίηση των παρατηρήσεων που περιγράφηκαν αναλυτικά στο κεφ. 4.2.1 και τους οποίους χρησιμοποιεί καθώς επεξεργάζεται, γραμμή προς γραμμή, κάθε ένα από τα RawInfoboxProperties files χωριστά. Αυτοί οι κανόνες βοηθούν τον **PDM** αφ ενός να αναγνωρίζει πότε μια γραμμή του αρχείου που επεξεργάζεται κάθε φορά αφορά πρόσωπο και αφ ετέρου να προσδιορίζει, όταν μπορεί, σε ποια κλάση της Wikipedia οντολογίας (αλλιώς κατηγορία) υπάγεται το πρόσωπο αυτό. Αυτό οδηγεί στη δημιουργία ενός πίνακα με τα πρόσωπα και τις κατηγορίες στις οποίες ανήκουν.

Μετά συμβουλευόμενος τον πίνακα που δημιούργησε, απομονώνει από κάθε αρχείο μιας έκδοσης τις γραμμές που περιγράφουν τα πρόσωπα και τις επεξεργάζεται. Κατά την επεξεργασία αυτή ο **PDM** αντλεί το όνομα και τις ιδιότητες κάθε οντότητας, απομονώνοντας συγκεκριμένες ακολουθίες χαρακτήρων στην κάθε γραμμή των αρχείων που διαβάζει (βλ. 4.2.1 του παρόντος) και την ακατέργαστη τιμή κάθε ιδιότητας, απομονώνοντας

ο,τιδήποτε ακολουθεί τους χαρακτήρες που περιγράφουν ιδιότητα. Τις πληροφορίες αυτές κατευθύνει προς έναν προσωρινό πίνακα, για την κάθε έκδοση, τον οποίο δημιουργεί στον οποίο και εισάγει τα ονόματα, τις ιδιότητες των οντοτήτων και τις αντίστοιχες τιμές τους, όπως αυτά εμφανίζονται σε κάθε συγκεκριμένη έκδοση, ή ισοδύναμα τις οντότητες με τα χαρακτηριστικά τους ή ισοδύναμα τα στιγμιότυπα των οντοτήτων κατά τα οριζόμενα στο κεφ.3.1 του παρόντος. (βλ. 4.2.3)

2. **Το component EC (Entities Creator)** που δημιουργεί την **Temporal Entities' Data Base (EDB)** (βλ. κεφ. 4.2.4).

Για τη δημιουργία της «χρονικά ευαίσθητης» Βάσης Δεδομένων Οντοτήτων, ο **EC** εκμεταλλεύεται τους προσωρινούς πίνακες, που δημιουργεί ο **PDM** για κάθε έκδοση.

Αξιοποιώντας τις στήλες ent_name, ent_category & prop_name των προσωρινών πινάκων κάθε έκδοσης, δημιουργεί αρχικά τους πίνακες persons, category & properties αντίστοιχα, που περιέχουν μια row για κάθε διακριτό όνομα οντότητας, κατηγορίας και ιδιότητας παράγοντας (autoincrement) και τα αντίστοιχα μοναδικά αντίστοιχα αναγνωριστικά (unique ids) τους.

Από το συνδυασμό των στηλών ent_name και ent_category δημιουργεί τον πίνακα που αποδίδει την ομαδοποίηση των προσώπων σε κατηγορίες. Σε κάθε row αυτού του πίνακα εισάγεται το μοναδικό αναγνωριστικό του προσώπου και το μοναδικό αναγνωριστικό της κατηγορίας στην οποία ανήκει, όπως προκύπτουν από την αξιοποίηση των πινάκων persons και category, που δημιουργήθηκαν προηγουμένως.

Από το συνδυασμό των στηλών ent_name και prop_name δημιουργεί:

α) τον πίνακα που αποδίδει τις ιδιότητες των προσώπων (perprop). Για καθεμιά από τις ιδιότητες που περιγράφουν ένα πρόσωπο εισάγεται σε αυτό τον πίνακα μια row με στήλες το μοναδικό αναγνωριστικό του προσώπου και το μοναδικό αναγνωριστικό της ιδιότητας, όπως αυτά προκύπτουν από την αξιοποίηση των πινάκων persons και properties, που δημιουργήθηκαν προηγουμένως.

β) τον πίνακα που αποδίδει την πληθυκότητα (v3XXX_howmany, βλ.κεφ.3.2 σελ.13) των χαρακτηριστικών κάθε προσώπου σε μια συγκεκριμένη έκδοση. Για καθεμιά από τις ιδιότητες που περιγράφουν ένα πρόσωπο εισάγεται σε αυτό τον πίνακα μια row με στήλες το μοναδικό αναγνωριστικό του προσώπου και το μοναδικό αναγνωριστικό της ιδιότητας, όπως προηγουμένως, καθώς και το πόσες φορές αποδίδεται κάποια τιμή στην ιδιότητα αυτή.

Τέλος δημιουργεί για κάθε έκδοση έναν πίνακα (v3XXX_perpropval) του οποίου κάθε γραμμή αντιστοιχεί και σε ένα χαρακτηριστικό προσώπου, αξιοποιώντας τις τιμές των στηλών ent_name, prop_name και raw_val του αντίστοιχου σε αυτήν προσωρινού πίνακα. Συγκεκριμένα

κάθε γραμμή αυτού του πίνακα περιέχει το μοναδικό αναγνωριστικό του προσώπου, της ιδιότητας (από τους πίνακες persons και properties) καθώς και την τιμή, όπως ακριβώς ανασύρθηκε από τα αρχικά Raw InfoBox Properties files.

3. **Το component Web Application,** που δίνει τη δυνατότητα στο χρήστη (άνθρωπο ή άλλο σύστημα) δύο δυνατότητες:

a) να ενεργοποιεί τον PDM, αφού πρώτα εκτελέσει κάποιες προπαρασκευαστικές εργασίες (π.χ καθορισμός διαδρομής εγκατάστασης ΒΔ, καθορισμός διαδρομής τοπικής αποθήκευσης των Raw Infobox Properties files), και
b) να απευθύνει στην TEDB ερωτήματα (queries) και να λαμβάνει αποτελέσματα από αυτήν.

Για την υλοποίηση των παραπάνω χρησιμοποιείται μια απλή διεπαφή (interface) με τους χρήστες (ανθρώπους), η οποία και υποδέχεται τις επιλογές τους (μελετητικές προτιμήσεις εκδόσεων ή/και προσώπων), τις ενσωματώνει κατάλληλα σε sql statements, τα οποία αποστέλλει στην **TEDB** όπου εκτελούνται και παράγουν τα αποτελέσματά τους. Στη συνέχεια λαμβάνει τα αποτελέσματα αυτά και τα σερβίρει ως HTML, σελίδες αλλά και ως αρχείο Excel.

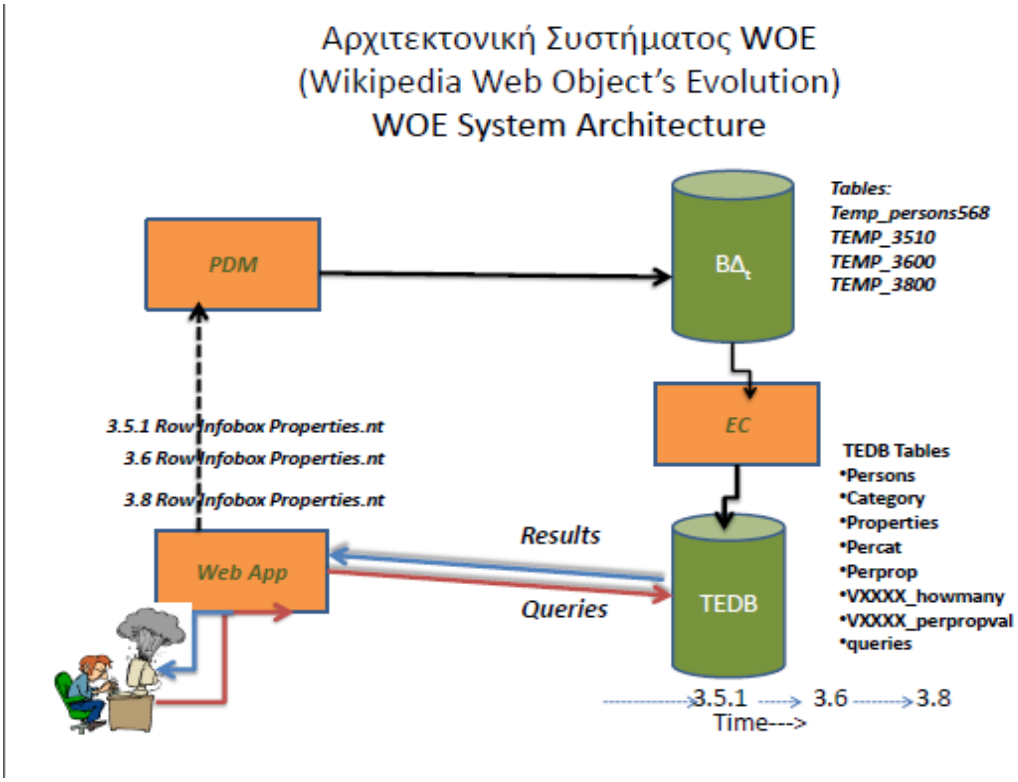
Συνοψίζοντας:

Το **component Persons' Data miner (PDM)** δέχεται ως είσοδο τα downloaded Raw InfoBox Properties files και παραδίδει μια σχεσιακή **$ΒΔ_t$** (την Data Base της εικόνας 5.1) .

Το **component Entities Creator (EC)** δέχεται ως είσοδο την σχεσιακή $ΒΔ_t$ και παραδίδει **την Temporal Entities' DataBase-TEDB** .

Το **component Web App** χρησιμοποιεί την **TEDB** και δίνει την δυνατότητα στο χρήστη του συστήματος (ο οποίος μπορεί να είναι ένας χρήστης Internet ή/και κάποιο άλλο σύστημα) να θέτει ένα κάθε φορά από τα προτεινόμενα ερευνητικά ερωτήματα (βλ. κεφ. 3.3) και να μελετά/αξιοποιεί τα αποτελέσματα που η **TEDB** επιστρέφει για το συγκεκριμένο ερώτημα.

Στην εικόνα 5.1, που ακολουθεί, φαίνεται σχηματικά η Αρχιτεκτονική αυτή.



Εικ.5.1 Η Αρχιτεκτονική του συστήματος WOE

Επισημαίνεται ότι το βέλος του χρόνου στην εικ. 5.1 δηλώνει τη χρονική διάσταση της δημιουργημένης Βάσης Δεδομένων των οντοτήτων (TEDB). Όπως έχει αναλυτικά περιγραφεί στο κεφ. 4.1, καθώς οι εκδόσεις της Wikipedia αλλάζουν, η πληροφορία για τα χαρακτηριστικά ενός Προσώπου αφορά μια συγκεκριμένη χρονική περίοδο, αυτήν που οριοθετείται από δύο διαδοχικά Wikipedia dumps. Έχοντας εξορύξει τα χαρακτηριστικά των προσώπων σε κάθε έκδοση, στην ουσία έχουμε καταγράψει την εξέλιξη των προσώπων στο χρόνο.

5.2 Προσφερόμενες Λειτουργίες

Το σύστημα WOE έχει αναπτυχθεί ως client server application με thin client. Επομένως ο χρήστης/client δεν έχει παρά να εισάγει την κατάλληλη IP στον browser του για να εισέλθει σε αυτό. Στο παράρτημα Α περιγράφονται τα βήματα που πρέπει να εκτελέσει κάποιος, προκειμένου να το εγκαταστήσει και να το λειτουργήσει τοπικά στον localhost:8080 (python server)

Το σύστημα WOE προσφέρει τη δυνατότητα στους χρήστες του να μελετήσουν την εξέλιξη των προσώπων που εμφανίζονται στην Wikipedia σε διάφορες χρονικές περιόδους, οι οποίες οριοθετούνται από τις αναθεωρήσεις των εκδόσεών της (versions), μέσω ερωτημάτων των

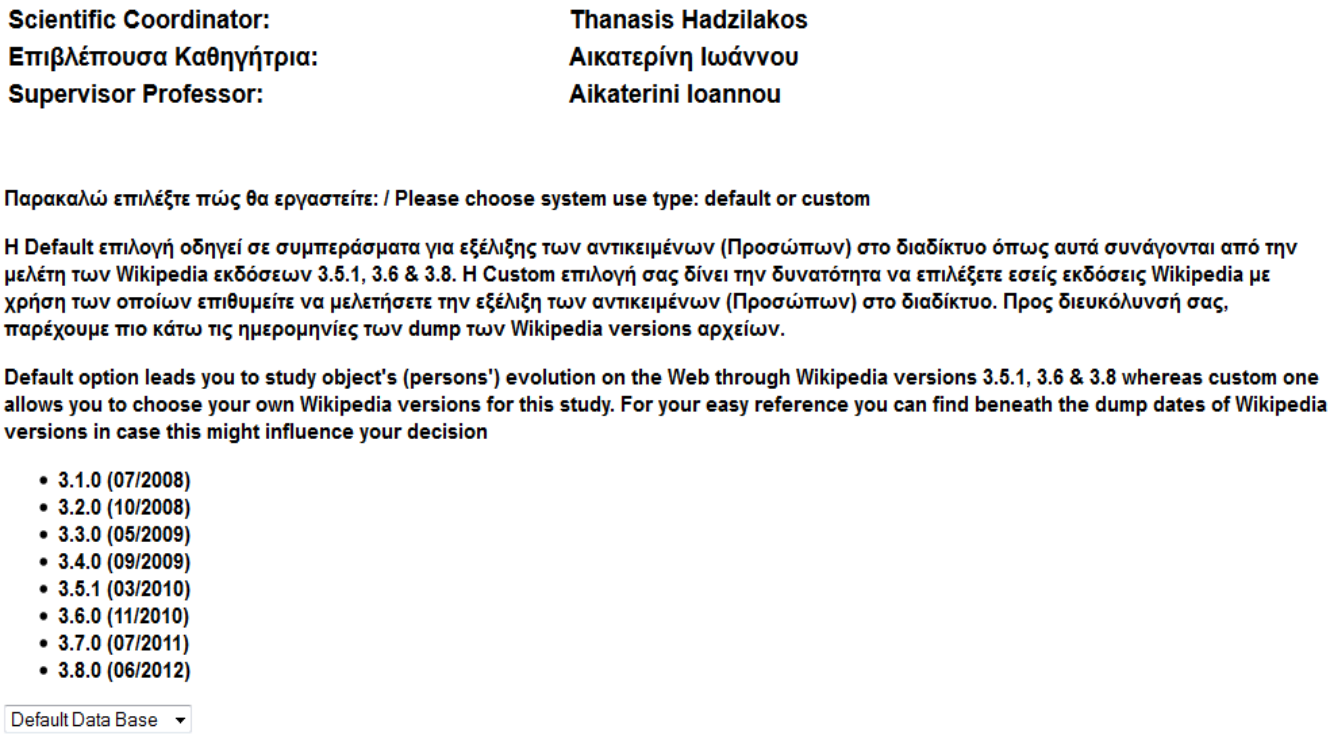
οποίων η δομή και η χρησιμότητα έχουν περιγραφεί αναλυτικά στο κεφάλαιο 3 της παρούσας μεταπτυχιακής διατριβής. Η δυνατότητα αυτή προσφέρεται σε 2 γλώσσες: Ελληνικά ή Αγγλικά. Το σύστημα WOE προσφέρει 2 βασικές λειτουργίες.

1. Λειτουργία αρχικοποίησης περιβάλλοντος και δημιουργία της **TEDB**
2. Ερευνητική Λειτουργία: Με αυτή ο χρήστης (άνθρωπος ή/και άλλο σύστημα) θέτει τα προτεινόμενα ερευνητικά ερωτήματα στην **TEDB** και λαμβάνει αποτελέσματα προς περαιτέρω αξιοποίηση.

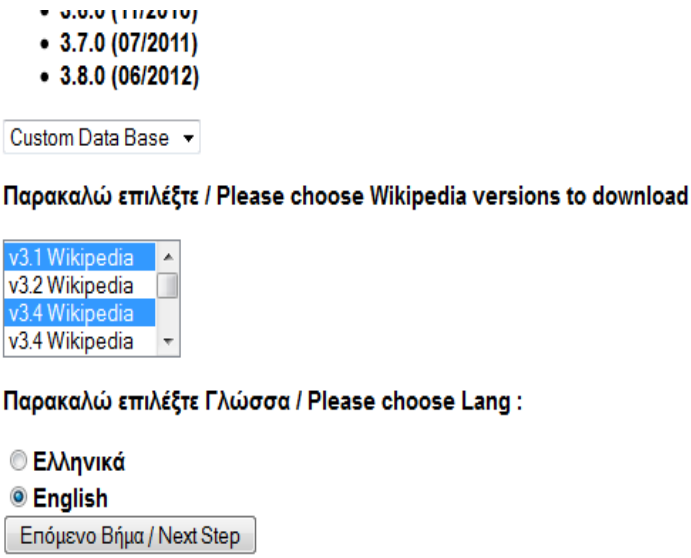
Με την είσοδό του στο WOE ο χρήστης καλείται να επιλέξει ποια από τις δύο μεθόδους δουλειάς προτιμά, όπως φαίνεται στην παρακάτω εικόνα 5.2, αφού πρώτα λάβει μια σύντομη ενημέρωση για τις δυνατές επιλογές που αυτό παρέχει.

5.2.1 Αρχικοποίηση Περιβάλλοντος – Δημιουργία TEDB

Από τη στιγμή που ο χρήστης επιλέξει την custom εγκατάσταση, καλείται να διαλέξει τις Wikipedia εκδόσεις με τις οποίες επιθυμεί να εργαστεί (Εικ. 5.3) και των οποίων τα Raw InfoBox Properties files πρέπει κάνει download.

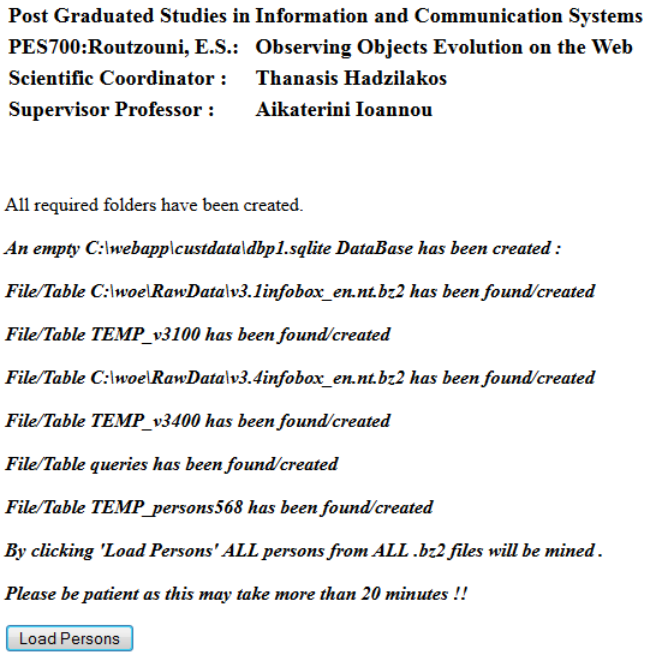


Εικόνα 5.2 Επιλογή τρόπου εργασίας στην αρχική σελίδα του συστήματος WOE



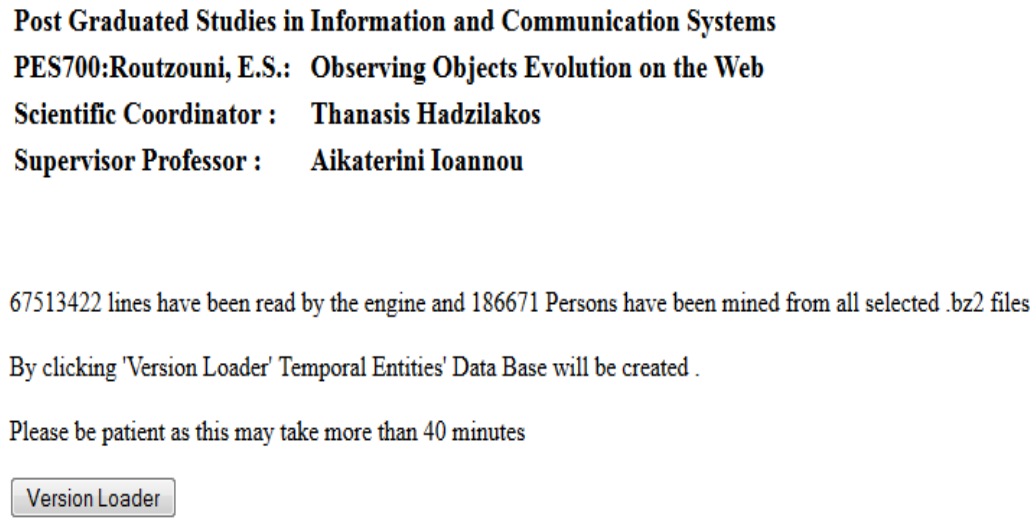
Εικόνα 5.3 Επιλογή εκδόσεων για dowload των αρχείων Raw Infobox Properties (.bz2 files), καθώς και γλώσσας εργασίας όπου έχουν επιλεγεί οι εκδόσεις 3.1 και 3.4 και η Αγγλική ως γλώσσα εργασίας

Στην συνέχεια το σύστημα εκτελεί το download των αντίστοιχων Raw Infobox Properties αρχείων και εκτελεί μια σειρά από προπαρασκευαστικές εργασίες/ ρυθμίσεις που θα εξασφαλίσουν την ορθή λειτουργία του PDM component. Επιστρέφει ενημερωτικά μηνύματα (Εικ. 5.4) στο χρήστη (για "ψυχολογικούς λόγους ", μια και δεν απαιτείται κάποια δράση εκ μέρους του, παρά η συγκατάθεσή του για την εκκίνηση του PDM, αλλά και για πρακτικούς ώστε να μην εκκινήσει το PDM, αν κάτι δεν δουλέψει κατά τα αναμενόμενα (π.χ broken link).



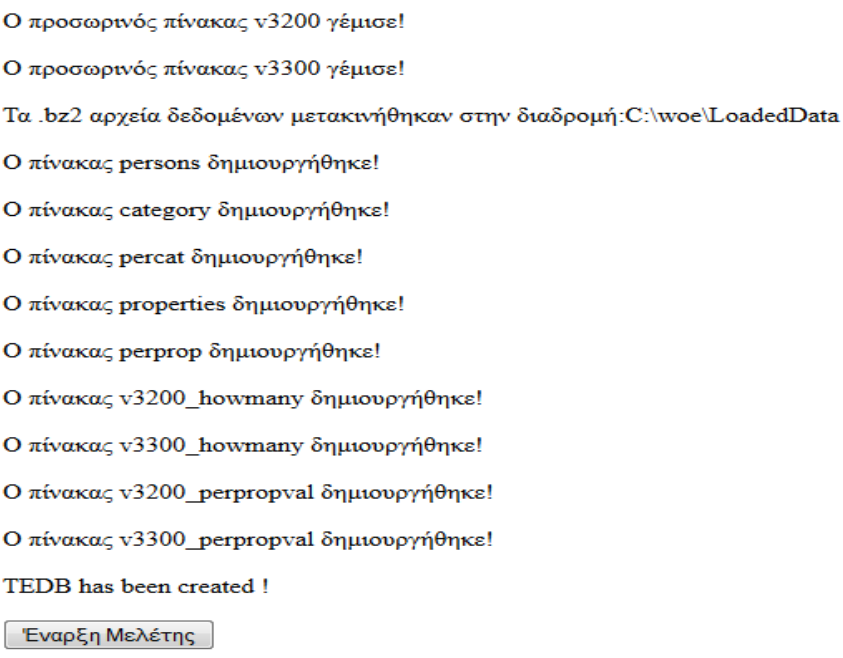
Εικόνα 5.4 Ενημερωτικά μηνύματα κατά την 1η φάση της custom χρήσης του συστήματος WOE.

Πατώντας ο χρήστης το Load Person, ξεκινά ο PDM που εξορύσσει τα πρόσωπα από τα Raw Infobox Properties αρχεία κάθε επιλεγμένης έκδοσης. Στο τέλος της εξόρυξης, ο χρήστης λαμβάνει σχετικό ενημερωτικό μήνυμα και καλείται να προχωρήσει στη δημιουργία της TEDB, όπως φαίνεται στην εικόνα 5.5, όπου δηλώνεται ότι 186671 πρόσωπα εξορύχθηκαν.



Εικόνα 5.5 Τα αποτελέσματα του PDM component

Στη συνέχεια το EC component αναλαμβάνει να δημιουργήσει την TEDB, όπως αναλυτικά έχει περιγραφεί σε προηγούμενα εδάφια του παρόντος (βλ.4.4, 5.1), επιστρέφοντας σχετικό μήνυμα με το πέρας των εργασιών του και καλώντας το χρήστη να εισέλθει στη Μελετητική Λειτουργία. (Εικ. 5.6)



Εικόνα 5.6 Τα αποτελέσματα του EC component

5.2.2 Ερευνητική Λειτουργία.

Η ερευνητική λειτουργία που παρέχεται από το σύστημα WOE συνίσταται σε υποβολή των ερευνητικών ερωτημάτων (queries), που έχουν περιγραφεί αναλυτικά στο κεφ.3.3 της παρούσας μεταπτυχιακής διατριβής, στην TEDB προς λήψη αποτελεσμάτων.

Επειδή:

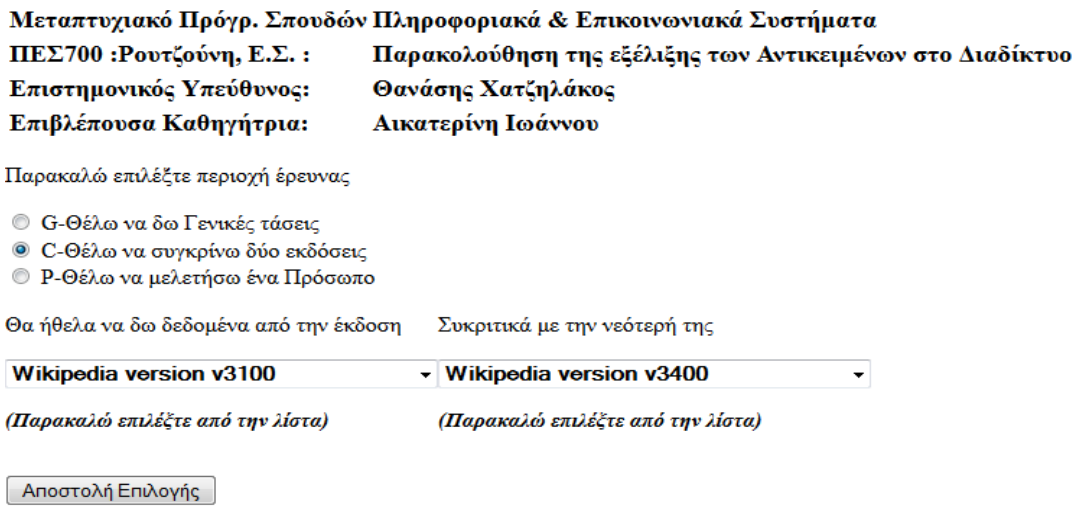
α) οι συγκρίσεις γίνονται πάντα ανά δύο εκδόσεις της Wikipedia και

β) στην default εγκατάσταση έχουμε 3 εκδόσεις παρούσες την 3.5.1, 3.5 και 3.8 και

γ) στην custom εγκατάσταση μπορεί να έχουμε πάνω από 2 (αλλά τουλάχιστον 2) εκδόσεις

αμέσως μετά την είσοδό σε Ερευνητική Λειτουργία, πρέπει να επιλεγεί περιοχή εργασίας. Να επιλεγεί, δηλαδή, αν θα μελετηθούν δεδομένα από μια έκδοση της Wikipedia, ή θα συγκριθούν δύο Wikipedia εκδόσεις μεταξύ τους ή θα αναζητηθούν δεδομένα για ένα συγκεκριμένο πρόσωπο), αλλά και η/οι Wikipedia έκδοση/εκδόσεις στις οποίες θα εστιαστεί η μελέτη όπως φαίνεται στην εικόνα 5.7

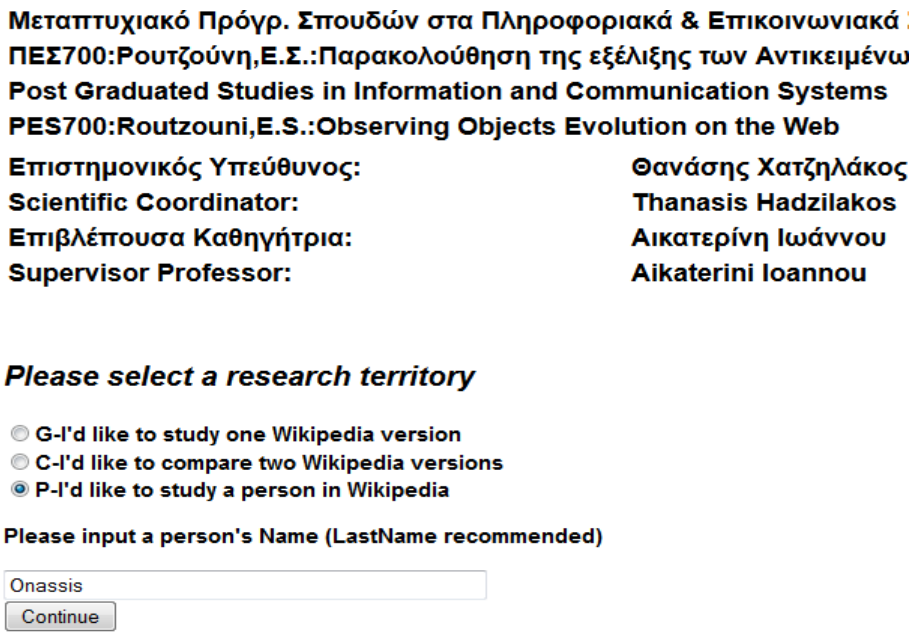
Η καθεμιά από τις τρεις επιλογές οδηγεί σε μια ομάδα από ερευνητικά ερωτήματα από τα οποία μόνο ένα μπορεί να επιλεχθεί κάθε φορά. Στην εικόνα 5.7 έχει επιλεγεί να συγκριθούν οι Wikipedia εκδόσεις 3.1 και 3.4 των οποίων τα δεδομένα μπήκαν στο σύστημα σε προηγούμενο στάδιο.



Εικόνα 5.7 Επιλογή ερευνητικής περιοχής και συγκεκριμένων Wikipedia εκδόσεων για την παρακολούθηση της εξέλιξης των οντοτήτων σε αυτές.

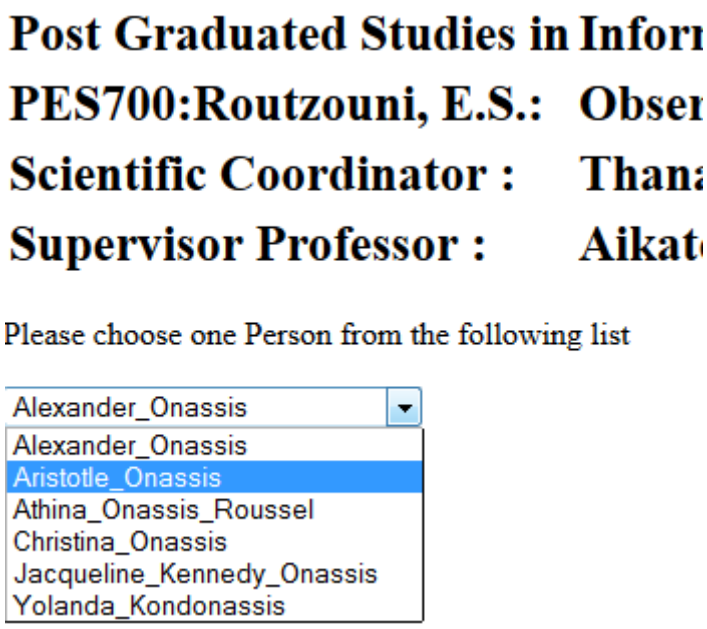
Για να μελετηθεί ένα πρόσωπο, θα πρέπει πρώτα να αναζητηθεί. Το σύστημα δίνει τη δυνατότητα, όπως φαίνεται στην παρακάτω εικόνα 5.8, να εισαχθεί το όνομα του προσώπου προς εντοπισμό, στα Αγγλικά, αφού χρησιμοποιούμε Αγγλικές εκδόσεις της Wikipedia.

Στην εικόνα 5.8 αναζητάται κάποιο πρόσωπο στου οποίου το όνομα περιέχεται το "Onassis".

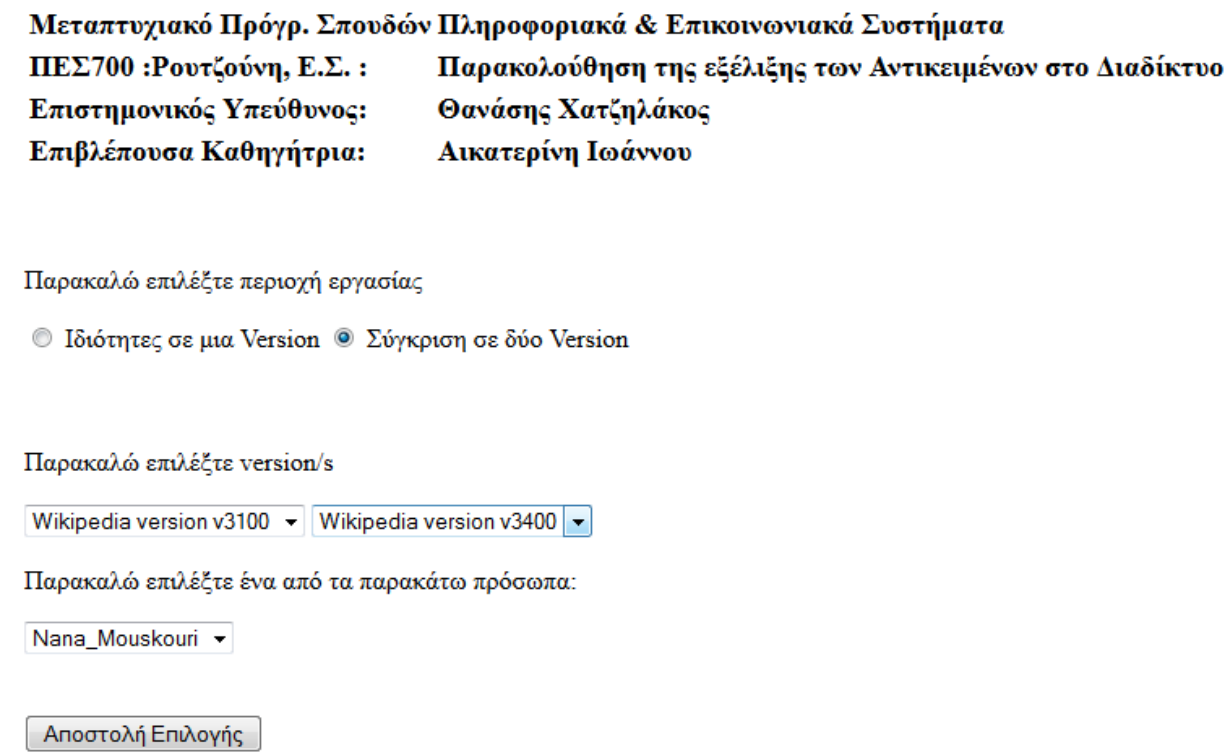


Εικόνα 5.8 Δυνατότητα Αναζήτησης του Προσώπου προς μελέτη.

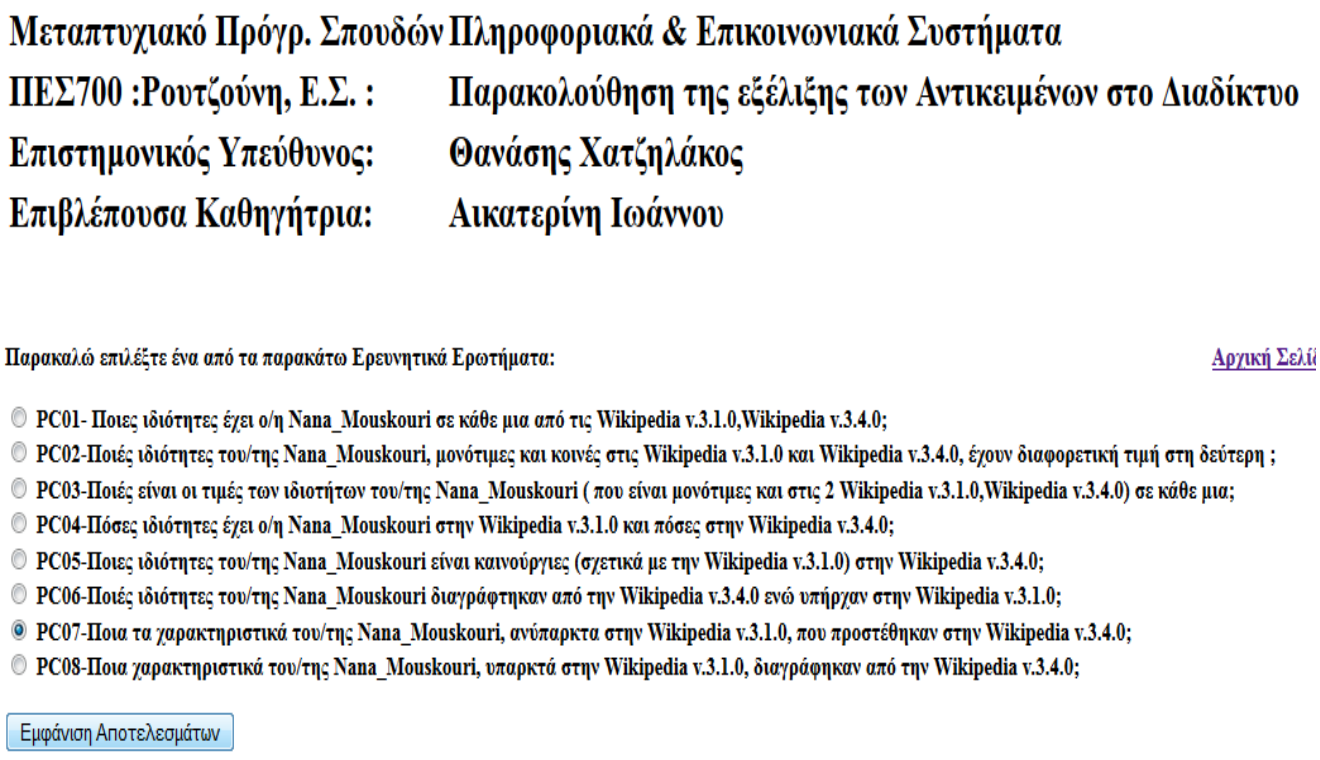
Στην εικόνα 5.9, που ακολουθεί, πρέπει να επιλεχθεί το πρόσωπο προς μελέτη ανάμεσα στα πρόσωπα που επιστράφηκαν από το σύστημα και στο όνομα τους περιέχεται το "Onassis".
Τα φωτογραφικά στιγμιότυπα 5.8 και 5.9 προέρχονται από την default εγκατάσταση.
Στην εικόνα 5.10 βλέπουμε ένα φωτογραφικό στιγμιότυπο που προέρχεται από custom use του συστήματος WOE για τις εκδόσεις 3.1 και 3.4 στο οποίο έχει επιλεγεί να μελετηθεί η εξέλιξη του προσώπου "Nana_Mouskouri" στις εκδόσεις 3.1 και 3.4.



Εικόνα 5.9 Επιλογή του προσώπου Aristotle_Onassis για μελέτη της εξέλιξής του ανάμεσα στα πολλά πρόσωπα που επιστράφηκαν και φέρουν αυτό ή παρόμοιο επώνυμο.

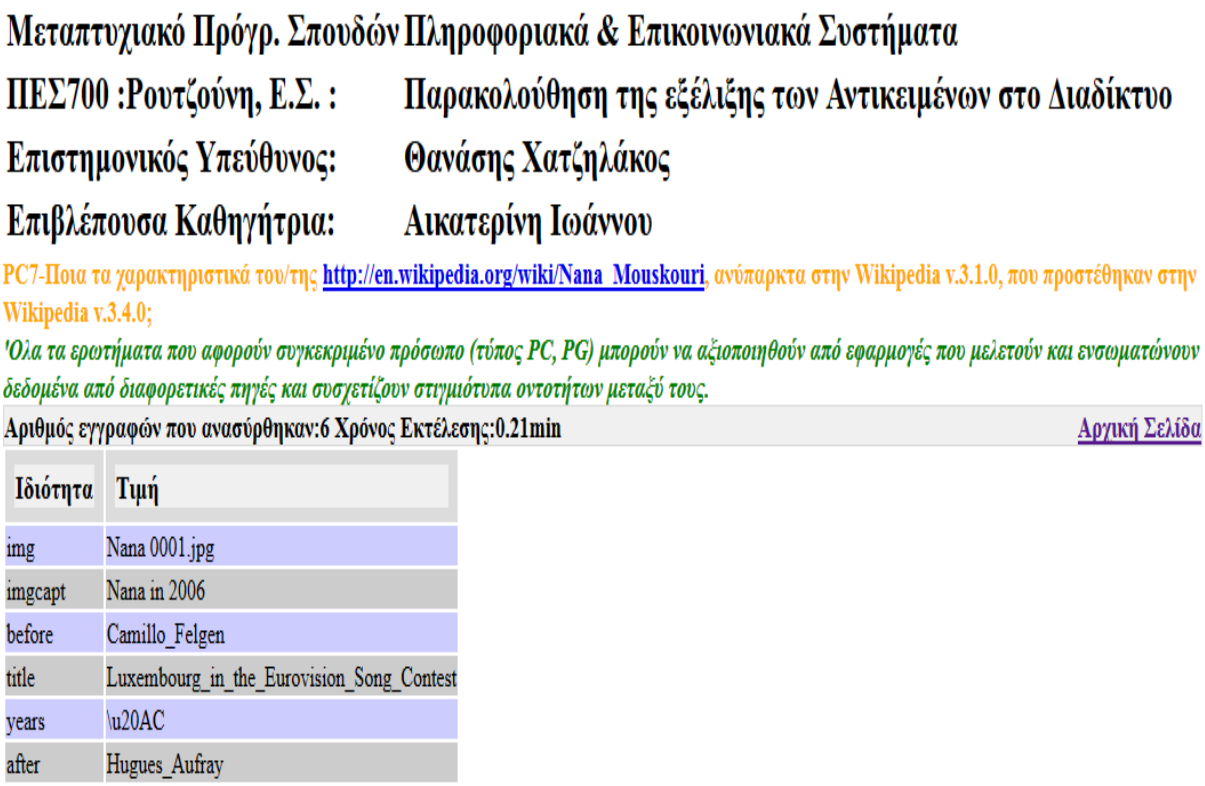


Εικόνα 5.10 Επιλογή της παρακολούθησης της εξέλιξης του προσώπου "Nana_Mouskouri" στις εκδόσεις Wikipedia3.1 και 3.4



Εικόνα 5.11 Ερευνητικά ερωτήματα που συνδέονται με την παρακολούθηση της εξέλιξης του προσώπου "Nana_Mouskouri" στις Wikipedia εκδόσεις 3.1 και 3.4

Στην εικόνα 5.11 έχει επιλεγεί να διαπιστώσουμε ποια χαρακτηριστικά της Νάνας Μούσχουρη είναι καινούργια στην έκδοση 3.4. Τα αποτελέσματα αυτού του ερωτήματος φαίνονται στην εικόνα 5.12 που ακολουθεί.



Μεταπτυχιακό Πρόγρ. Σπουδών Πληροφοριακά & Επικοινωνιακά Συστήματα

ΠΕΣ700 :Ρουτζούνη, Ε.Σ. : Παρακολούθηση της εξέλιξης των Αντικειμένων στο Διαδίκτυο

Επιστημονικός Υπεύθυνος: Θανάσης Χατζηλάκος

Επιβλέπουσα Καθηγήτρια: Αικατερίνη Ιωάννου

PC7-Ποια τα χαρακτηριστικά του/της http://en.wikipedia.org/wiki/Nana_Mouskouri, ανύπαρκτα στην Wikipedia v.3.1.0, που προστέθηκαν στην Wikipedia v.3.4.0;

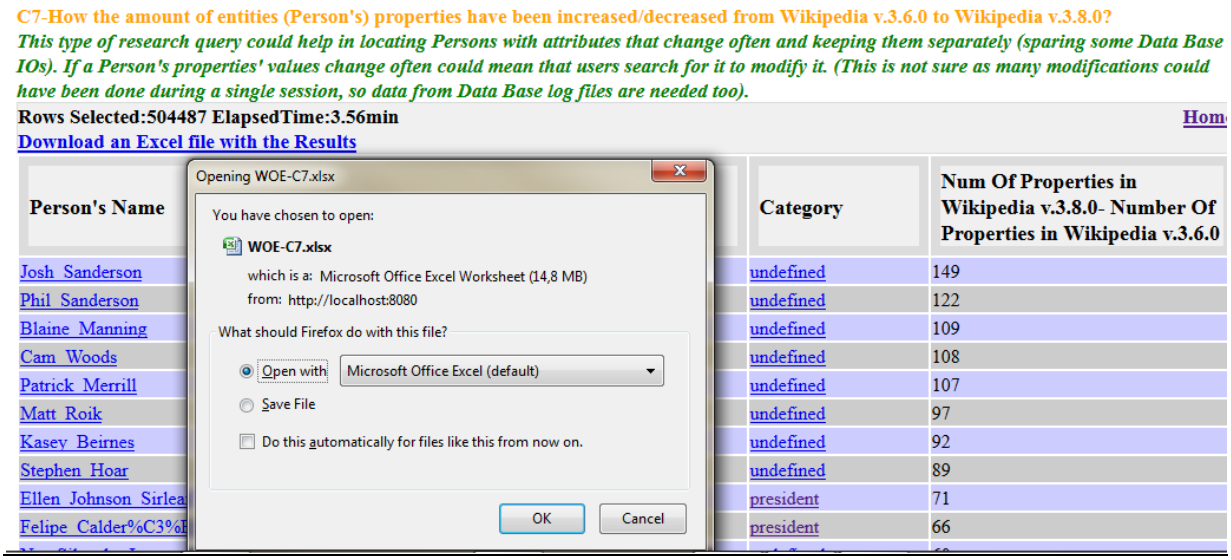
'Ολα τα ερωτήματα που αφορούν συγκεκριμένο πρόσωπο (τύπος PC, PG) μπορούν να αξιοποιηθούν από εφαρμογές που μελετούν και ενσωματώνουν δεδομένα από διαφορετικές πηγές και συσχετίζουν στιγμιότυπα οντοτήτων μεταξύ τους.

Αριθμός εγγραφών που ανασύρθηκαν:6 Χρόνος Εκτέλεσης:0.21min Αρχική Σελίδα

Ιδιότητα	Τιμή
img	Nana 0001.jpg
imgcapt	Nana in 2006
before	Camillo_Felgen
title	Luxembourg_in_the_Eurovision_Song_Contest
years	\u20AC
after	Hugues_Aufray

Εικόνα 5.12 Τα χαρακτηριστικά της Νάνας Μούσχουρη που προστέθηκαν στην Wikipedia 3.4

Το σύστημα δίνει τη δυνατότητα download όλων των αποτελεσμάτων των ερευνητικών ερωτημάτων σε μορφή Excel αρχείου, όπως φαίνεται, για παράδειγμα, στην εικόνα 5.13 που ακολουθεί.



Εικ. 5.13 Δυνατότητα λήψης των αποτελεσμάτων σε μορφή XL.

5.3 Μελλοντικές Επεκτάσεις

Το σύστημα WOE που υλοποιήσαμε επιδέχεται αρκετές επεκτάσεις μεταξύ των οποίων ξεχωρίζουμε:

1. Την on the fly επέκταση των ερευνητικών ερωτημάτων με ερωτήματα που εξυπηρετούν τον χρήστη και τα οποία θα εισάγονται από αυτόν λίγο πριν την εκτέλεσή τους.

2. Τη δυνατότητα επιλογής μιας TEDB, ανάμεσα σε πολλές διαφορετικές TEDBs που ενδεχομένως θα δημιουργηθούν στον server από custom χρήσεις, σε διαφορετικές χρονικές στιγμές. Προς το παρόν το σύστημα WOE "βλέπει" την TEDB που δημιουργείται ΜΟΝΟ ως το output μιας συγκερκιμένης custom χρήσης του, δηλ. για να εισέλθει στη λειτουργία των ερευνητικών ερωτημάτων, κατά την διάρκεια μιας custom χρήσης, πρέπει υποχρεωτικά να περάσει από τη φάση της αρχικοποίησης περιβάλλοντος που περιγράφηκε στο κεφ.5.2.1 του παρόντος.

- Το 1 θεωρούμε ότι αποτελεί σημαντικότατη επέκταση του συστήματος WOE διότι θα το καταστήσει εργαλείο πιο ευέλικτο, πιο αποδοτικό και πιθανά δημοφιλέστερο. Πρέπει να τονιστεί ότι το σύστημα WOE, χάρη στην υλοποίηση που επιλέχθηκε με τον πίνακα queries, είναι απόλυτα επεκτάσιμο. Θεωρητικά άπειρα ερευνητικά ερωτήματα μπορούν να εισέλθουν άμεσα στο σύστημα, ακόμα και για την default εγκατάσταση, αρκεί να καταχωριστούν τα αντίστοιχα SQLs στον πίνακα queries, ακολουθώντας την προτυποποίηση, που έχει ήδη εν συντομία περιγραφεί στο κεφ.3 και περιγράφεται αναλυτικότερα στο παράρτημα Γ1. Στην επέκταση που προτείνουμε, δεν αναφερόμαστε στην προσθήκη ερευνητικού ερωτήματος που κάποιος θα μας παράσχει σε νεκρό για το σύστημα χρόνο και εμείς θα το εντάξουμε σε αυτό, αλλά για τη δυνατότητα του συστήματος να αυτο-επεκτείνεται on the fly κατά τη διάρκεια της χρήσης του από κάποιον user (πρόσωπο ή σύστημα)
- Το 2 θεωρούμε ότι είναι σημαντικό, διότι μας απαλλάσσει από χρονοβόρες (πέραν της μιας ώρας) διαδικασίες της custom εγκατάστασης, που τώρα είναι απαραίτητες προκειμένου να δημιουργηθεί κάθε TEDB.

5.4 Σύνοψη Συμπερασμάτων

Από την παρούσα μεταπτυχιακή διατριβή προκύπτουν τα ακόλουθα:

Το dump των αρχείων της Wikipedia version 3.5.1 έλαβε χώρα τέλη Μαρτίου του 2010.

Το dump των αρχείων της Wikipedia version 3.6 έλαβε χώρα το Νοέμβριο του 2011, ενώ το dump των αρχείων της Wikipedia version 3.8 έλαβε χώρα τέλη Ιουνίου του 2012.

Αντίστοιχα το πλήθος των προσώπων στην 3.5.1 είναι 430.379 , στην 3.6 φθάνει το 561.671 (μεταβολή +30,50%), ενώ στην 3.8 αγγίζουν το 1.023.177 ! (+82,17%). Εάν τώρα λάβουμε υπόψη μας το χρόνο στον οποίο η μεταβολή αυτή συντελέστηκε, θα δούμε ότι έχουμε στην πρώτη περίπτωση (από την 3.5.1 στην 3.6) ένα ρυθμό μεταβολής της τάξης των +6565 προσώπων /μήνα, ενώ στη δεύτερη περίπτωση (από την 3.6 στην 3.8) ένα ρυθμό μεταβολής της τάξης του +57688 πρόσωπα/μήνα.

Τα υπό μελέτη πρόσωπα μπορούν να ομαδοποιηθούν σε κατηγορίες, το πλήθος των οποίων παραμένει σχετικά σταθερό (58 στην 3.5.1 και 59 στις υπόλοιπες), αλλά παρατηρούμε ότι υπάρχουν κατηγορίες πολυπληθέστερες άλλων που δίνουν σαφείς ενδείξεις ότι θα πρέπει να έχουν αφοσιωμένους εξυπηρετητές. Τέτοιες κατηγορίες είναι οι κάτοχοι κυβερνητικών θέσεων (Wikipedia term:"office holders"), οι μουσικοί καλλιτέχνες (Wikipedia term:"musical_artist") ,οι παίκτες κρίκετ, οι στρατιωτικοί (Wikipedia term:"military_persons"), οι παίκτες χόκεϊ επί πάγου, οι συγγραφείς, οι επιστήμονες, οι καλλιτέχνες, οι χαρακτήρες comics (Wikipedia term:"comic_character") και οι κολεγιακοί προπονητές, (Wikipedia term:"college_coach") για να αναφέρουμε τις 10 πολυπληθέστερες, κατά φθίνουσα ταξινόμηση.

Οι μεταβολές του πλήθους των ιδιοτήτων ανά κατηγορία (της τάξεως κάποιων εκατοντάδων) δεν είναι τόσο εντυπωσιακές σε απόλυτους αριθμούς από έκδοση σε έκδοση. Τα σκήπτρα κρατούν πάλι τα κυβερνητικά στελέχη, οι μουσικοί, οι συγγραφείς , ηθοποιοί, οι καλλιτέχνες, οι δημιουργοί comics και οι επιστήμονες κατά φθίνουσα σειρά ταξινόμησης. Εν τούτοις θα πρέπει να σκεφτούμε ότι η αύξηση του πλήθος των ιδιοτήτων κατά κάποιες εκατοντάδες αν συνδυαστεί με την τεράστια αύξηση του πλήθους των προσώπων, οδηγεί σε εκθετική αύξηση των απαιτήσεων σε χώρους.

Στην τρέχουσα (κατά τη περίοδο της εκπόνησης της παρούσας μεταπτυχιακής διατριβής, 10/2012-05/2013) έκδοση Wikipedia (v.3.8) 5.572 ιδιότητες εμφανίζονται για πρώτη φορά σε ένα διάστημα 8 μηνών από την προγενέστερη της έκδοση (v.3.6).

Τα πρόσωπα που έχουν ιδιότητες με μια μόνο τιμή σε κάθε version, που όμως είναι διαφορετική στη νεότερη από ότι στη παλιότερη, είναι πάρα πολλά (γι'αυτό και για να μπορέσουμε να δούμε αποτελέσματα στην οθόνη μας σε λογικό χρόνο, ζητήσαμε τυχαία να μας επιστραφούν μόνο αυτά που έχουν pid modulo 11 = 0). Αν όμως παρατηρήσουμε με προσοχή τις τιμές αυτές που επιστρέφει το ερώτημα, θα δούμε ότι οι διαφορές είναι ένα '_' , ένας κενός χαρακτήρας, ένας προσδιορισμός που αποσαφηνίζει το χαρακτηριστικό, αλλά δεν το μεταβάλει στην ουσία του.

Ένα από τα κύρια θέματα που αναδύθηκε από τη μελέτη μας είναι αυτό της απόδοσης ιδιοτήτων σε πρόσωπα από τους χρήστες, με τρόπο ώστε να περιορίζονται οι πολλαπλές αναφορές. Θα

μπορούσε για παράδειγμα να αναπτυχθεί ένα framework για τους χρήστες της Wikipedia, που να τους διευκολύνει όταν θέλουν να εισάγουν μια ιδιότητα για ένα πρόσωπο. Θα μπορούσε, για παράδειγμα, να προηγείται μια αναζήτηση και να επιστρέφονται προτάσεις από τις οποίες ο χρήστης να μπορεί να επιλέγει και να εισάγει μια νέα ιδιότητα μόνο αν δεν βρίσκει κάποια αρκετά ικανοποιητική.

Τέλος, ενώ τα πρόσωπα που έχουν τα πιο πολλά χαρακτηριστικά ανά κατηγορία δεν φαίνεται να διατηρούνται σταθερά με την πάροδο του χρόνου, ο γνωστός μας Arnold_Schwarzenegger κρατά σταθερά τα σκήπτρα του ηθοποιού με τις περισσότερες ιδιότητες και ο Geroge H. W. Bush (Bush πατέρας) αυτά του προέδρου με τις περισσότερες ιδιότητες, αλλά και του προσώπου με τις περισσότερες ιδιότητες γενικότερα. Η συζήτηση για το αν το να συγκεντρώνει ένα πρόσωπο τις περισσότερες ιδιότητες στην κατηγορία του σημαίνει ότι είναι και το δημοφιλέστερο ξεφεύγει από τα όρια της παρούσας μεταπτυχιακής διατριβής.

Τέλος πρέπει να επισημανθεί ότι πολλά άλλα συμπεράσματα θα μπορούσαν να εξαχθούν με την χρήση του WOE που όμως εξαρτώνται από το σύστημα που μελετά τα δεδομένα.

Κεφάλαιο 6
Επίλογος - Συμπεράσματα

6.1 Ανακεφαλαίωση

Στα πλαίσια της μεταπτυχιακής διατριβής αυτής δημιουργήσαμε ένα σύστημα (WOE) το οποίο εξορύσσει, αναλύει και μελετά την εξέλιξη αντικειμένων που περιγράφουν οντότητες στον πραγματικό κόσμο.

Στο σύστημα αυτό εισάγαγαμε δεδομένα που περιγράφουν και μοντελοποιούν πρόσωπα, όπως αυτά εμφανίζονται στην Αγγλική έκδοση της Wikipedia (http://en.wikipedia.org/), σε διάφορες χρονικές περιόδους, που αντιστοιχούν στο χρονικό διάστημα μεταξύ των dump συγκεκριμένων εκδόσεών της (versions): της 3.5.1 (dumped 3/2010), της 3.6 (dumped 11/2011) και της 3.8 (dumped 6/2012), που ήταν και η τρέχουσα κατά το χρόνο εκπόνησης της μεταπτυχιακής διατριβής.

Έλαβε χώρα επεξεργασία των δεδομένων αυτών από τα διάφορα components του συστήματος και δημιουργήθηκε μια χρονική Βάση Δεδομένων Οντοτήτων / Temporal Entities' Data Base (TEDB), που παρουσιάζει τις ίδιες οντότητες σε διαφορετικές χρονικές περιόδους και άρα την εξέλιξή τους στο χρόνο (καινοτομία). Ερευνητικά ερωτήματα απευθύνθηκαν στη βάση αυτή των οποίων τα αποτελέσματα είναι αξιοποιήσιμα προς μελέτη από ανθρώπους ή/και άλλα συστήματα.

Μερικές από τις διαπιστώσεις:

- Εκθετικοί ρυθμοί μεταβολής σε πλήθος προσώπων και ιδιοτήτων (από 430.379 τον 3/2010 σε 1.023.177 τον 6/2012)

- Σταθερό σχετικά πλήθος κατηγοριών προσώπων (π.χ. Ηθοποιός, πολιτικός) (57 τον 11/11 , 58 τον 6/2012)
- Ισχυρές διαφοροποιήσεις στον αριθμό των προσώπων-μελών κάθε κατηγορίας που δίνουν σαφείς ενδείξεις ότι κάποιες από τις κατηγορίες αυτές θα πρέπει να τύχουν ειδικής μεταχείρισης κατά την διαχείρισή τους. Σταθερά στις 3 πρώτες θέσεις μουσικοί καλλιτέχνες με 60718 πρόσωπα τον 6ο/2012 έναντι 47503 τον 3ο/2010 , οι κάτοχοι κυβερνητικών θέσεων (39902-22135) και στρατιωτικοί (21787-14020).
- Ισχυρότερος ρυθμός αύξησης για πλήθος προσώπων και ιδιοτήτων στην κατηγορία "κάτοχοι κυβερνητικών θέσεων". (17767 περισσότερα πρόσωπα και 461 περισσότερες ιδιότητες από τον 3ο/2010 στον 6ο/2012)
- Μέτρια αύξηση του πλήθους των ιδιοτήτων κάθε κατηγορίας, που συνδυασμένη όμως με τον εκρηκτικό αριθμό αύξησης των μελών της αυξάνει ακόμα περισσότερο τις απαιτήσεις σε χώρους (461 περισσότερες για τους κατόχους κυβερνητικών θέσεων τον 6ο/2012 από τον 3ο/2010 και 440 για τους μουσικούς).
- Σημαντική είσοδος νέων ιδιοτήτων στο σύστημα. (6274 περισσότερες τον 6ο/2012 σε σχέση με τον 3ο/2010)
- Πολλαπλές αναφορές και ανάγκη προτυποποίησης του τρόπου απόδοσης ιδιοτήτων στις οντότητες (Πρόσωπα)
- Αλλαγές στα χαρακτηριστικά προσώπων που δεν είναι πάντα αλλαγές ουσίας, αλλά αφορούν σε εμφάνιση, (π.χ. ένα space λιγότερο, ένα underscore περισσότερο) ή αποσαφήνιση. (π.χ. “Smyrna” , “Smyrna, Otoman Empire”, για περισσότερα βλ. αποτελέσματα του C14 ερωτήματος).
- Πρόσωπα που είναι περισσότερο δημοφιλή από άλλα στην κατηγορίας τους, με την έννοια ότι έχουν μεγαλύτερο πλήθος ιδιοτήτων, η απόκτηση των οποίων υποδηλώνει ότι οι χρήστες ασχολήθηκαν μαζί τους περισσότερο
- Καταχωρίσεις που, ενώ δεν είναι πρόσωπα στον πραγματικό κόσμο, εμφανίζονται ως πρόσωπα στηνWikipedia. π.χ "presidency of Bill Clinton", "84th Delaware Assembly".

Όλα τα παραπάνω μπορούν να αξιοποιηθούν από αλγορίθμους και συστήματα διαχείρισης ΒΔ, προκειμένου να βελτιωθεί η συνολική εμπειρία του χρήστη.

6.2 Επεκτάσεις

Αντίστοιχα με την περίπτωση της Wikipedia, το σύστημα WOE θα μπορούσε να επεκταθεί να αποσπάσει οντότητες, ιδιότητες και χαρακτηριστικά με παρόμοιο τρόπο και από άλλες πηγές ευμετάβλητων δεδομένων.

Το σύστημα WOE θα μπορούσε να επεκταθεί και σε περισσότερο δομημένα δεδομένα, με σκοπό την παρακολούθηση της εξέλιξης των προσώπων, για τα οποία ενδιαφέρονται G2G services.

Έρευνες δείχνουν ότι οι χρήστες ΑΝΑΖΗΤΟΥΝ συγκεκριμένα γνωρίσματα των στιγμιοτύπων για τα οποία ενδιαφέρονται [1] και το σύστημα WOE μπορεί να φανεί πολύ χρήσιμο, καθώς εύκολα απαντά στο ερώτημα ποια ακριβώς είναι τα χαρακτηριστικά του στιγμιότυπου μιας οντότητας σε μια συγκεκριμένη χρονική περίοδο.

Επίσης η συνολική περιηγητική εμπειρία των χρηστών μπορεί να βελτιωθεί, αν αξιοποιηθούν τα ευρήματα του WOE που αφορούν στην δημοφιλία συγκεκριμένων κατηγοριών στιγμιοτύπων (π.χ. Μουσικοί καλλιτέχνες).

Το σύστημα WOE παρέχει επίσης πληροφορία για την κατάσταση των οντοτήτων στο χρόνο, η οποία μπορεί να φανεί χρήσιμη σε εφαρμογές που αντλούν χώρο-χρονικές πληροφορίες (βλ. [5]), καθώς ιδιότητες που έχουν σχέση με τοπικά δεδομένα και αφορούν ένα πρόσωπο σε μια χρονική περίοδο (αυτή που προηγήθηκε του dump των Wikipedia versions) είναι πολύ εύκολο να εντοπιστούν με το σύστημα WOE.

Κεφάλαιο 7

Βιβλιογραφία

[01] Nilesh N. Dalvi, Ravi Kumar, Bo Pang, Raghu Ramakrishnan, Andrew Tomkins, Philip Bohannon, Sathiya Keerthi, Srujana Merugu: A web of concepts. PODS 2009: 1-12.

[02] Alon Y. Halevy, Michael J. Franklin, David Maier: Principles of dataspace systems. PODS 2006: 1-9

[03] Ekaterini Ioannou, Wolfgang Nejdl, Claudia Niederée, Yannis Velegrakis: On-the-Fly Entity-Aware Query Processing in the Presence of Linkage. PVLDB 3(1): 429-438 (2010)

[04] Flavio Rizzolo, Yannis Velegrakis, John Mylopoulos, Siarhei Bykau: Modeling Concept Evolution: A Historical Perspective. ER 2009: 331-345

[05] Yafang Wang, Bin Yang, Spyros Zoupanos, Marc Spaniol, Gerhard Weikum: Scalable spatio-temporal knowledge harvesting. WWW (Companion Volume) 2011: 143-144

Παράρτημα Α

Α1.Περιβάλλον Εργασίας

Εργαστήκαμε σε PC Lenovo με λειτουργικό σύστημα Windows 7 Home Premium 64-bit, επεξεργαστή Intel Core i3-2330M CPU @2.20 GHz και RAM 6 GB

Α2.Οδηγίες εγκατάστασης

Το σύστημα WOE (Wikipedia Web Object's Evolution) παρέχει δύο δυνατότητες χρήσης:

1. **τη default,** όπου ο χρήστης χρησιμοποιεί τη ΒΔ των προσώπων όπως προέκυψε από την επεξεργασία των .nt files στα πλαίσια της μεταπτυχιακής διατριβής αυτής.

2. **Την custom**, όπου ο χρήστης του συστήματος WOE καλείται να επιλέξει τα RawInfoboxProperties.bz2 files των Wikipedia versions που αυτός επιθυμεί να επεξεργαστεί, να δημιουργήσει την αντίστοιχη ΒΔΠ, να θέσει σε αυτήν τα προτεινόμενα ερωτήματα και να λάβει τα αντίστοιχα αποτελέσματα.
Και για τις 2 περιπτώσεις ο χρήστης εισέρχεται στο σύστημα μέσω του web browser Mozilla. δίνοντας την IP του server που παρέχει την υπηρεσία.

Εάν, εν τούτοις, κάποιος χρήστης θέλει να εγκαταστήσει και να εκτελεί το σύστημα τοπικά στο localhost τότε πρέπει*:*

1. Από τον σύνδεσμο http://python.org/download/ να κατεβάσει και να εγκαταστήσει την γλώσσα προγραμματισμού Python v.3.3 (Windows) χωρίς να αλλάξει τα προτεινόμενα.

2. Από τον σύνδεσμο https://addons.mozilla.org/el/firefox/addon/sqlite-manager/ να κατεβάσει και να εγκαταστήσει το sqlite manager plug-in για τον Firefox, προαιρετικά, αν θέλει να "βλέπει¨ την βάση μεμονωμένα.

3. Από το συνοδευτικό CD/.zip file να αντιγράψει ακριβώς κάτω από τον root C: τους φακέλλους webapp και woe με τα περιεχόμενά τους.

4. Από το συνοδευτικό CD/.zip file να αντιγράψει τον φάκελλο openpyxl στην διαδρομή δίσκου C:\Python33\Lib (δημιουργείται στο βήμα 1)

5. Από την windows command line να εκτελέσει κατά σειρά τις ακόλουθες εντολές:
   A. C:>cd webapp
   το σύστημα απαντά:
   C:\webapp>
   B. C:\webapp>c:\Python33\python.exe simple_httpd.py

   Η τελευταία αυτή εντολή αυτή ξεκινά τον server και απαντά:
   Starting simple_httpd on port:8080

6. Στη γραμμή διευθύνσεων Firefox Browser να γράψει:
   http://locahost:8080/index.html.
   Το σύστημα WOE ξεκινά και η αρχική σελίδα του εμφανίζεται στον χρήστη.

Να επισημάνουμε ότι για την custom χρήση χρειάζεται να υπάρχει ενεργή σύνδεση στο διαδίκτυο προκειμένου να εκτελεστούν τα απαραίτητα file's downloads.

Παράρτημα Β

Β1.SQLs δημιουργίας πινάκων και εισαγωγής δεδομένων

Τα παρακάτω SQLs εκτελέστηκαν στην ΒΔ και παρήγαγαν τους πίνακες της Entities' Data Base dbpediapersons.sqlite

1. CREATE TABLE TEMP_persons568(name TEXT UNIQUE NOT NULL, category TEXT)
   Ο πίνακας γέμισε με κώδικα Python (findPersons.py)

2. CREATE TABLE TEMP_$ (rowcount INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL , ent_name TEXT,ent_category TEXT, prop_name TEXT,raw_val TEXT,fine_val TEXT)
   όπου $ v3510,v3600,v3810. Οι πίνακες γέμισαν με κώδικα Python (versionLoader.py)

3. CREATE TABLE persons(pid INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL, pname TEXT NOT NULL, plink TEXT)

4. CREATE TABLE category(cid INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL, cname TEXT NOT NULL, clink TEXT)

5. CREATE TABLE percat(percatid INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL,pid INTEGER NOT NULL , cid INTEGER NOT NULL,

FOREIGN KEY (pid) REFERENCES persons(pid) ON DELETE CASCADE ON UPDATE CASCADE,

FOREIGN KEY (cid) REFERENCES category(cid) ON DELETE CASCADE ON UPDATE CASCADE)

6. INSERT INTO CATEGORY(CNAME) SELECT DISTINCT lower(category) FROM TEMP_persons568 και
7. UPDATE CATEGORY SET CLINK= ""||philosopher||""

8. INSERT INTO PERSONS (PNAME) SELECT DISTINCT NAME FROM TEMP_persons568 και
9. update persons set plink = ""||pname||""

10. INSERT INTO percat (pid,cid)

select p.pid,c.cid from TEMP_persons568w join persons p on p.pname=w.name

join category c on c.cname=lower(w.category)

==========

ΔΗΜΙΟΥΡΓΙΑ ΚΑΙ Γ Ε Μ Ι Σ Μ Α Τ Ο Υ Π Ι Ν Α Κ Α Ρ Ο Ρ Ε Ρ Τ Ι Ε Σ

---

11. CREATE TABLE properties(propid INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL, propname TEXT NOT NULL)

ο properties γέμισε ως εξής:

********ολες οι properties της 3.5.1 (v3510_properties)

a) create table propT1 as

SELECT DISTINCT lower(prop_name) pname from TEMP_3510

********ολες οι properties της 3.6.0 (v3600_properties)

b) create table propT2 as

SELECT DISTINCT lower(prop_name) pname from TEMP_3600

********ολς οι properties της 3.8 (v3800_properties)

c) create table propT3 as

SELECT DISTINCT lower(prop_name) pname from TEMP_3800

**********ΕΝΑ UNION ΠΑΝΩ ΤΟΥΣ

d) create table propT3 as SELECT * FROM propT1 UNION SELECT * FROM propT2

********** και γεμιζει ο πίνακας properties

12. INSERT INTO properties (propname) SELECT prop_name from propT3

********** D R O P οι TEMPORARY propT1,T2,T3

e) Drop table propT1
f) Drop table propT2
g) Drop table propT3

==========

ΔΗΜΙΟΥΡΓΙΑ ΚΑΙ ΓΕΜΙΣΜΑ ΤΟΥ ΠΙΝΑΚΑ ΙΔΙΟΚΤΗΣΙΑΣ ΑΝΤΙΚΕΙΜΕΝΟΥ ΙΔΙΟΤΗΤΑΣ

Η ΛΟΓΙΚΗ ΕΙΝΑΙ ΟΛΕΣ ΟΙ ΔΥΝΑΤΕς ΙΔΙΟΚΤΗΣΙΕΣ ΝΑ ΑΠΕΙΚΟΝΙΣΤΟΥΝ ΣΕ ΕΝΑ ΠΙΝΑΚΑ

ΑΝ ΜΙΑ VERSION ΣΥΜΜΕΤΕΧΕΙ ΣΤΗΝ ΙΔΙΟΚΤΗΣΙΑ ΑΥΤΟ ΣΗΜΑΙΝΕΙ ΟΤΙ Η ΣΧΕΣΗ ΤΗΣ ΘΑ ΕΧΕΙ

ΕΓΓΡΑΦΗ ΣΤΟΝ ΠΙΝΑΚΑ ΤΙΜΩΝ (DETAIL)

==========

----------Δ Η Μ Ι Ο Υ Ρ Γ Ε Ι Τ Α Ι Ο Π Ι Ν Α Κ Α Σ ΣΧΕΣΗ ΙΔΙΟΚΤΗΣΙΑΣ

13. CREATE TABLE perprop(perpropid INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL,pid INTEGER NOT NULL , propid INTEGER NOT NULL,
FOREIGN KEY (pid) REFERENCES persons(pid) ON DELETE CASCADE ON UPDATE CASCADE,
FOREIGN KEY (propid) REFERENCES properties(propid) ON DELETE CASCADE ON UPDATE CASCADE)

--ΣΤΗΝ ουσία πρόκειται για όλες τις σχέσεις που έχουν υπάρξει μέχρι σήμερα στην wikipedia persons αφού η 3.8 είναι η latest version

----------βρίσκονται οι σχέσεις από τον TEMP_3800 ΜΕ ΤΗΝ ΠΛΗΘΥΚΟΤΗΤΑ ΤΟΥΣ ΚΑΙ ΓΡΑΦΟΝΤΑΙ ΣΕ ΠΡΟΣΩΡΙΝΟ

a) create table t1 as

SELECT ent_name, lower(prop_name) propname , count(*) howmany from TEMP_3800 group by ent_name, lower(prop_name)

Από αυτόν τον προσωρινό δημιουργείται ένας άλλος που εχει μόνο την σχέση με ids και οχι ονόματα και ΧΩΡΙΣ ΠΛΗΘΥΚΟΤΗΤΑ

b) create table t2 as

SELECT p.pid,prop.propid from t1 join persons p on t1.ent_name=p. pname join properties prop on t1.propname=prop.propname

---------- βρίσκονται οι σχέσεις από τον TEMP_3600 ΜΕ ΤΗΝ ΠΛΗΘΥΚΟΤΗΤΑ ΤΟΥς ΚΑΙ ΓΡΑΦΟΝΤΑΙ ΣΕ ΠΡΟΣΩΡΙΝΟ

c) create table t3 as

SELECT ent_name, lower(prop_name) propname , count(*) howmany from TEMP_3600 group by ent_name, lower(prop_name)

Από αυτόν τον προσωρινό δημιουργείται ένας άλλος που εχει μόνο την σχέση με ids και οχι ονόματα και ΧΩΡΙΣ ΠΛΗΘΥΚΟΤΗΤΑ

d) create table t4 as

SELECT p.pid,prop.propid from t3 join persons p on t3.ent_name=p. pname join properties prop on t3.propname=prop.propname

---------- βρίσκονται οι σχέσεις από τον TEMP_3510 ΜΕ ΤΗΝ ΠΛΗΘΥΚΟΤΗΤΑ ΤΟΥς ΚΑΙ ΓΡΑΦΟΝΤΑΙ ΣΕ ΠΡΟΣΩΡΙΝΟ

e) create table t5 as

SELECT ent_name, lower(prop_name) propname , count(*) howmany from TEMP_3510 group by ent_name, lower(prop_name)

Από αυτόν τον προσωρινό δημιουργείται ένας άλλος που εχει μόνο την σχέση με ids και οχι ονόματα και ΧΩΡΙΣ ΠΛΗΘΥΚΟΤΗΤΑ

f) create table t6 as

SELECT p.pid,prop.propid from t5 join persons p on t5.ent_name=p. pname join properties prop on t5.propname=prop.propname

----------Ενώνονται οι πίνακες σχέσεων ιδιοκτησίας από τις 3 versions ΜΕ UNION, οι όμοιες εγγραφές θα απαλειφθούν.

g) create table t7 as

SELECT * from t2 union select * from t4 union select * from t6

-------Τ Ω Ρ Α Ε Ι Μ Α Σ Τ Ε Ε Τ Ο Ι Μ Ο Ι Ν Α εισάγουμε στον πινακα perprop

14. INSERT INTO perprop (pid,propid) select pid,propid from t5 order by pid,propid

----------DROP OI t2,t4,t6 save for later oi t1,t3,t5

h) Drop table t2
i) Drop table t4
j) Drop table t6

==========

Δ Η Μ Ι Ο Υ Ρ Γ Ο Υ Ν Τ Α Ι Ο Ι Π Ι Ν Α Κ Ε Σ Π Λ Η Θ Υ Κ Ο Τ Η Τ Α Σ Γ Ι Α Κ Α Θ Ε Ε Κ Δ Ο Σ Η

==========

15. CREATE TABLE v3510_howmany(perpropid INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL, howmany integer,
FOREIGN KEY (perpropid) REFERENCES perprop(perpropid) ON DELETE CASCADE ON UPDATE CASCADE)

16. CREATE TABLE v3600_howmany(perpropid INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL, howmany integer,
FOREIGN KEY (perpropid) REFERENCES perprop(perpropid) ON DELETE CASCADE ON UPDATE CASCADE)

17. CREATE TABLE v3800_howmany(perpropid INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL, howmany integer,
FOREIGN KEY (perpropid) REFERENCES perprop(perpropid) ON DELETE CASCADE ON UPDATE CASCADE)

----------Τ Ω Ρ Α ΘΑ ΕΙΣΑΓΟΥΜΕ ΣΤΟΥΣ ΠΙΝΑΚΕΣ ΤΗΣ ΠΛΗΘΥΚΟΤΗΤΑΣ :

---------- για την πληθυκότητα της v3800

είναι ΗΔΗ ΕΤΟΙΜΟΣ Ο t1

a) create table t2 as select p.pid, prop.propid,t1.howmany from
t1 join persons p on t1.ent_name = p.pname join properties prop on t1.propname = prop.propname

------ join perprop για perpropid

b) create table t30 as select p.perpropid, t2.howmany from t2 join perprop p on
p.pid=t2.pid and p.propid=t2.propid

---------- για την πληθυκότητα της v3600
είναι ΗΔΗ ΕΤΟΙΜΟΣ Ο t3

c) create table t4 as select p.pid, prop.propid,t3.howmany from t3 join persons p on t3.ent_name = p.pname join properties prop on t3.propname = prop.propname

------ join perprop για perpropid

d) create table t40 as select p.perpropid, t4.howmany from t4 join perprop p on p.pid=t4.pid and p.propid=t4.propid

---------- για την πληθυκότητα της v3600
είναι ΗΔΗ ΕΤΟΙΜΟΣ Ο t5

e) create table t6 as select p.pid, prop.propid,t5.howmany from t5 join persons p on t5.ent_name = p.pname join properties prop on t5.propname = prop.propname

------ join perprop για perpropid

f) create table t50 as select p.perpropid, t6.howmany from t6 join perprop p on
p.pid=t6.pid and p.propid=t6.propid

Γεμίζουν οι πίνακες
insert into v3800_howmany(perpropid,howmany) select * from t30
insert into v3600_howmany(perpropid,howmany) select * from t40
insert into v3510_howmany(perpropid,howmany) select * from t50

----------D R O P TEMPORARY propT1,T2,T3,t4,t5,t6,t30,t40,t50

g) Drop table propT1, Drop table propT1, κτλ

==========

ΤΕΛΟΣ ΦΤΙΑΧΝΟΝΤΑΙ ΟΙ ΠΙΝΑΚΕΣ ΜΕ ΤΑ DETAILS οπου σε καθε σχέση ιδιοκτησίας perprop, που υπάρχει στην version αντιστοιχίζουν μια τιμή

18. create table v3510_perpropval(rowcount INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL,perpropid INTEGER, rawval TEXT,fineval TEXT,
FOREIGN KEY (perpropid) REFERENCES perprop(perpropid) ON DELETE CASCADE ON UPDATE CASCADE)

19. create table v3600_perpropval(rowcount INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL,perpropid INTEGER, rawval TEXT,fineval TEXT,
FOREIGN KEY (perpropid) REFERENCES perprop(perpropid) ON DELETE CASCADE ON UPDATE CASCADE)

20. create table v3800_perpropval(rowcount INTEGER PRIMARY KEY AUTOINCREMENT UNIQUE NOT NULL,perpropid INTEGER, rawval TEXT,fineval TEXT,
FOREIGN KEY (perpropid) REFERENCES perprop(perpropid) ON DELETE CASCADE ON UPDATE CASCADE)

****

a) create table temp3510_detail as select pp.perpropid,t.raw_val,"" fineval
from TEMP_3510 t join persons p on p.pname =t.ent_name
join properties prop on prop.propname=lower(t.prop_name)
join perprop pp on p.pid =pp.pid and prop.propid = pp.propid

21. insert into v3510_perpropval(perpropid,rawval,fineval) select perpropid,raw_val,fineval from temp3510_detail

---------->>>>>>>>>>>>>>>

b) create table temp3600_detail as select pp.perpropid,t.raw_val,"" fineval
from TEMP_3600 t join persons p on p.pname =t.ent_name join properties prop
on prop.propname=lower(t.prop_name)
join perprop pp on p.pid =pp.pid and prop.propid = pp.propid

22. insert into v3600_perpropval(perpropid,rawval,fineval) select perpropid,raw_val,fineval from temp3600_detail

---------->>>>>>>>>>>>>>>

c) create table temp3800_detail as select pp.perpropid,t.raw_val,"" fineval
from TEMP_3800 t join persons p on p.pname =t.ent_name
join properties prop on prop.propname=lower(t.prop_name)
join perprop pp on p.pid =pp.pid and prop.propid = pp.propid

23. insert into v3800_perpropval(perpropid,rawval,fineval) select perpropid,raw_val,fineval from temp3800_detail

==========

ΤΕΛΟΣ ΦΤΙΑΧΝΕΤΑΙ Ο ΠΙΝΑΚΑΣ queries

24. CREATE TABLE queries (qcat CHAR NOT NULL , qid INTEGER NOT NULL , qsql TEXT, qdesc_el TEXT, qdesc_en TEXT, outputTableHeaders_el TEXT, outputTableHeaders_en TEXT, quse_el TEXT, quse_en TEXT, PRIMARY KEY (qcat , qid))

B2.SQLs εξόρυξης τιμών ιδιοτήτων

Τα παρακάτω SQLs «καθαρίζουν» τις τιμές των ιδιοτήτων, ώστε να εμφανίζονται στον χρήστη σε κατανοητή μορφή.

1. update vXXXX_perpropval set fineval =REPLACE(rawval ,"<http://dbpedia.org/resource/","")
2. update vXXXX _perpropval set fineval =REPLACE(fineval ,"> .","")
3. update vXXXX _perpropval set fineval =REPLACE(fineval ,"||@en .","")
4. update vXXXX _perpropval set fineval =REPLACE(fineval ,"||^^<http://www.w3.org/2001/XMLSchema#int","")
5. update vXXXX _perpropval set fineval =REPLACE(fineval ,"||^^<http://www.w3.org/2001/XMLSchema#date","")
6. update vXXXX _perpropval set fineval =REPLACE(fineval ,"||^^<http://www.w3.org/2001/XMLSchema#gMonthDay","")
7. update vXXXX _perpropval set fineval =ltrim(ltrim(fineval ,"|,"))
8. update vXXXX _perpropval set fineval =ltrim(ltrim(fineval ,"""))

όπου vXXXX = v3510, v3600, v3800

Παράρτημα Γ

Γ1. Εισαγωγή επιπλέον ερευνητικών ερωτημάτων

Για να εισαχθούν (και να λειτουργούν ορθά) επιπλέον ερευνητικά ερωτήματα στο πίνακα queries πρέπει να ακολουθηθούν οι εξής συμβάσεις:

1. εάν το ερώτημα είναι γενικό και δεν αφορά συγκεκριμένο πρόσωπο στη στήλη qcat πρέπει να έχει την τιμή G
2. εάν το ερώτημα είναι συγκριτικό και δεν αφορά συγκεκριμένο πρόσωπο στη στήλη qcat πρέπει να έχει την τιμή C
3. εάν το ερώτημα είναι γενικό και αφορά συγκεκριμένο πρόσωπο στη στήλη qcat πρέπει να έχει την τιμή PG
4. εάν το ερώτημα είναι συγκριτικό και αφορά συγκεκριμένο πρόσωπο στη στήλη qcat πρέπει να έχει την τιμή PC
5. Ο τίτλος του ερωτήματος που θέλουμε να εμφανιστεί στον χρήστη προς επιλογή πρέπει να καταχωριστεί στις στήλες qdesc-el και qdesc-en σε ελληνική και αγγλική γλώσσα αντίστοιχα.
6. το sql statement πρέπει να καταχωριστούν στη στήλη qsql τηρουμένων κάποιων συμβάσεων (βλ. και 8, 9 πιο κάτω)
7. οι επικεφαλίδες των στηλών αποτελεσμάτων που θέλουμε να εμφανιστούν στον χρήστη πρέπει να καταχωριστούν στις στήλες outputTableHeaders-el, outputTableHeaders-en

σε ελληνική και αγγλική γλώσσα αντίστοιχα χωρισμένες μεταξύ τους με «;» (semicolumn) (δες παράδειγμα 1)

8. στο όνομα του πίνακα, που στο sql αντιπροσωπεύει την παλαιότερη version, πρέπει να γραφτεί @, ενώ στο όνομα του πίνακα, που στο sql αντιπροσωπεύει την παλαιότερη version, πρέπει να γραφτεί #.

Παράδειγμα 1: έστω το sql που επιστρέφει ποιες ιδιότητες μιας έκδοσης (επιλογής του χρήστη) ΔΕΝ υπήρχαν σε κάποια ΣΥΓΚΕΚΡΙΜΕΝΗ προηγούμενη (επίσης επιλογής του χρήστη):

```
Select   distinct pp.propid,prop.propname
from     perprop pp
join     #_howmany hm on     pp.perpropid = hm.perpropid
join     properties prop on      prop.propid=pp.propid
where not exists
        (select pp1.propid from perprop pp1 join @_howmany hm1 on pp1.perpropid = hm1.perpropid
                where  pp.propid = pp1.propid)
```

Επειδή ΔΕΝ γνωρίζουμε ποιες θα είναι οι επιλογές του χρήστη, εισάγουμε το @ ως placeholder για την παλαιότερη και το # για την νεότερη έκδοση αντίστοιχα . Αυτονόητο είναι ότι εάν στο sql μας υπεισέρχεται μια μόνο version (τύπου G ή PG), τότε χρησιμοποιούμε μόνο το **@.**
(βλ.παράδειγμα 2)
Οι επικεφαλίδες Ελληνικές (Αγγλικές) στην στήλη outputTableHeaders-el (outputTableHeaders-en) πρέπει να είναι ως εξής:

Κωδικός Ιδιότητας**;**Όνομα Ιδιότητας

9. εάν στο sql αφορά σε μελέτη συγκεκριμένου προσώπου τότε το όνομα του προσώπου παριστάνεται με ?.

Παράδειγμα 2: έστω το sql που επιστρέφει το πλήθος ιδιοτήτων ενός προσώπου (επιλογής του χρήστη) σε κάποια (μία) έκδοση (επίσης επιλογής του χρήστη)

```
select w.pid,p.pname Person , c.cname Category, w.howManyPropsInPers  howManyPropsInPers
from persons p join
 (select pp.pid, count(pp.propid) howManyPropsInPers from  perprop pp join @_howmany hm
on hm.perpropid = pp.perpropid group by pp.pid order by howManyPropsInPers) w
on p.pid=w.pid join percat pc on pc.pid = w.pid join category c on c.cid=pc.cid
where p.pname = "?"
```

Επειδή ΔΕ γνωρίζουμε ποιες θα είναι οι επιλογές του χρήστη, εισάγουμε το @ ως placeholder για την έκδοση και το ? για το όνομα προσώπου αντίστοιχα .
Ειδικά για τους τίτλους των ερευνητικών ερωτημάτων χρησιμοποιείται ως placeholder, αντί προσώπου, το $ προς αποφυγή σύγχισης με το Αγγλικό ερωτηματικό.