ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ. ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

«Συλλογή, ολοκλήρωση και ανάλυση δεδομένων για Covid-19 με τεχνολογίες σημασιολογικού ιστού»

Πατσιμάς Ανδρέας

Α.Μ.: me2033

Επιβλέπων καθηγητής: κ. Δουλκερίδης Χρήστος

Φεβρουάριος 2022

Περίληψη

Η παρούσα διπλωματική εργασία αποσκοπεί στον συνδυασμό ετερογενών δεδομένων για την εύρεση συσχετίσεων μεταξύ μεταβλητών και την εξαγωγή συμπερασμάτων σχετικά με την εξέλιξη της πανδημίας Covid-19. Αρχικά συλλέχθηκαν δεδομένα από επίσημες πηγές που αναφέρονται άμεσα στις μετρικές της πανδημίας, δεδομένα που αφορούν την ατμοσφαιρική μόλυνση και δεδομένα σχετικά με τις μετακινήσεις του πληθυσμού. Τα δεδομένα αυτά οργανώθηκαν και δομήθηκαν σύμφωνα με τις αρχές του σημασιολογικού ιστού, ενώ δημιουργήθηκε και η αντίστοιχη οντολογία με τη χρήση της γλώσσας OWL και του εργαλείου Protégé. Στη συνέχεια με τη δημιουργία κατάλληλων λεξικών και γραμματικών κανόνων και τη χρήση RDF-Gen παράχθηκαν τα τελικά αρχεία που περιέχουν τις επιθυμητές RDF τριπλέτες. Έτσι η εξαγωγή των επιθυμητών πληροφοριών πραγματοποιήθηκε με την κατασκευή των κατάλληλων SPARQL ερωτημάτων που βασίστηκαν στη δομή της οντολογίας αυτής. Επιπρόσθετα, με τη χρήση της SPARQL και αντίστοιχων queries συγκεντρώθηκαν δευτερεύοντα δεδομένα που προέρχονται από πηγές του σημασιολογικού ιστού και σχετίζονται με τον αριθμό των νοσοκομείων, των αθλητικών εγκαταστάσεων και της πυκνότητας του πληθυσμού για ένα σύνολο πρωτευουσών συγκεκριμένων χωρών-πιλότων. Σύμφωνα με τα αποτελέσματα, φαίνεται να υπάρχουν ενδιαφέρουσες συσχετίσεις ανάμεσα στα δεδομένα που χρήζουν περεταίρω διερεύνησης.

Abstract

The present dissertation aims to combine heterogeneous data to find correlations between variables and to draw conclusions about the evolution of the Covid-19 pandemic. Initially, data were collected from officially sources that directly refer to pandemic metrics, data on air pollution and data on population mobility. This data were organized and structured according to the principles of semantic web, while the corresponding ontology was created using the OWL language and the Protégé tool. Then, with the creation of appropriate lexical and grammatical rules and the use of RDF-Gen, the final files were produced that include the desired RDF triplets. Therefore, the extraction of the desired information was carried out by constructing the appropriate SPARQL queries, which were based on the ontology's structure. In addition, secondary data was collected using SPARQL and corresponding queries stemming from their semantic web sources which are related to the number of hospitals, sports facilities, and population density for a set of specific pilot country capitals. According to the results, it seems that exist interesting correlations between the data that need further investigation.

Ευχαριστίες

Με την παρούσα διπλωματική εργασία ολοκληρώνονται οι σπουδές μου στο μεταπτυχιακό πρόγραμμα σπουδών «Μεγάλα Δεδομένα και Αναλυτική» του Τμήματος Ψηφιακών Συστημάτων στο Πανεπιστήμιο Πειραιά.

Στις σπουδές μου ήταν καθοριστική η συμβολή των καθηγητών μου στα γνωστικά αντικείμενα που παρακολούθησα, στους οποίους οφείλω να εκφράσω τις ειλικρινείς μου ευχαριστίες για τη συμβολή τους στην ολοκλήρωση των σπουδών μου.

Ιδιαίτερα επιθυμώ να ευχαριστήσω τον καθηγητή μου και επιβλέποντα στην παρούσα διπλωματική εργασία, κύριο Δουλκερίδη Χρήστο, για την επιστημονική και συμβουλευτική καθοδήγηση που μου προσέφερε σε όλα τα στάδια εκπόνησης της εργασίας με τις εύστοχες και πολύ εποικοδομητικές παρατηρήσεις του.

Τέλος οφείλω να εκφράσω τις ευχαριστίες μου προς τον ερευνητή κύριο Σαντιπαντάκη, χωρίς τη βοήθεια του οποίου δε θα ήταν δυνατή η ολοκλήρωση της διεξαγωγή της έρευνας.

Περιεχόμενα

Λίστα εικόνων

Λίστα πινάκων

1. Εισαγωγή

Παρ' όλο που πολλά κράτη, οργανισμοί και ιδιωτικοί φορείς δημοσιεύουν δεδομένα σχετικά με Covid-19, όπως αριθμός κρουσμάτων ανά γεωγραφική περιοχή ή ηλικιακή ομάδα, τα δεδομένα αυτά παρέχονται με διαφορετικό σχήμα, τύπους και μορφότυπους, γενονός που καθιστά δύσκολη την ανάλυσή τους. Επιπλέον, τα δεδομένα αυτά συνήθως δε διαθέτουν σαφή σημασιολογία και απαιτούν την ανθρώπινη παρέμβαση για την ερμηνεία και διασύνδεσή τους Ο σημασιολογικός ιστός, ο οποίος αποτελεί μία επέκταση του σημερινού παγκόσμιου ιστού, επιτρέπει την επεξεργασία και κατανόηση των δεδομένων, όχι μόνο από ανθρώπους, αλλά και από μηχανές. Αυτό προσφέρει τη δυνατότητα ενοποίησης δεδομένων από διαφορετικές πηγές και αξιοποίηση όλης αυτής της πληροφορίας.

Σκοπός της εργασίας είναι η συλλογή, ολοκλήρωση και ανάλυση δεδομένων σχετικών με Covid-19 από διάφορες πηγές, και η συσχέτισή τους με σκοπό την από κοινού ανάλυση δεδομένων που μπορεί να οδηγήσει στην ανακάλυψη χρήσιμων συμπερασμάτων και γνώσης.

Στην παρούσα εργασία συνδυάζονται δεδομένα από επίσημες πηγές με τεχνολογίες του σημασιολογικού ιστού, προκειμένου να δημιουργηθεί ένα νέο σύνολο δεδομένων, το οποίο αποσκοπεί στην εξαγωγή χρήσιμων συμπερασμάτων και γνώσης σχετικά με την πανδημία του Covid-19 και την εξάπλωσή της.

Η δομή της υπόλοιπης εργασίας είναι η ακόλουθη:

Στο κεφάλαιο 2 περιγράφονται βασικές έννοιες, όπως ο σημασιολογικός ιστός, το πλαίσιο περιγραφής πόρων – RDF, καθώς και το σχήμα RDFS, οντολογίες, κλπ, οι οποίες είναι απαραίτητες για την κατανόηση της παρούσας εργασίας.

Στο κεφάλαιο 3 περιγράφεται η υλοποίηση που αναπτύχθηκε στην παρούσα εργασία, ξεκινώντας από μία σύντομη επισκόπηση στις απαιτήσεις και τους στόχους, και συνεχίζοντας στη συλλογή και προετοιμασία των δεδομένων, στην ανάλυση του σχεδιασμού και ανάπτυξης της οντολογίας και τέλος, στο μετασχηματισμό των δεδομένων σε RDF.

Στο κεφάλαιο 4 παρουσιάζονται τα αποτελέσματα που προέκυψαν σε μορφή διαγραμμάτων και τα αντίστοιχα ερωτήματα – queries που χρησιμοποιήθηκαν για την εξαγωγή τους.

Τέλος, στο κεφάλαιο 5 αναλύονται τα συμπεράσματα που προέκυψαν από την παρούσα διπλωματική εργασία και προτείνονται μελλοντικές επεκτάσεις.

2 Θεωρητικό υπόβαθρο

2.1 Σημασιολογικός ιστός

Ο παγκόσμιος ιστός (World Wide Web ή www) αποτελεί ένα ανοιχτό σύστημα πολυμεσικού περιεχομένου και διασυνδεδεμένων πληροφοριών, το οποίο δίνει τη δυνατότητα στους χρήστες του διαδικτύου να αναζητούν πληροφορίες μεταβαίνοντας από ένα έγγραφο σε ένα άλλο [10].

Η αναζήτηση περιεχομένου συγκεκριμένης σημασίας αποτελεί έναν από τους συνηθέστερους λόγους χρήσης του διαδικτύου. Όμως, μία βασική του αδυναμία είναι η ανάγκη της ανθρώπινης παρέμβασης για τον εντοπισμό περιεχομένου με απορρέουσα σημασία.

Ο σημασιολογικός ιστός (Web 3.0 ή Semantic Web) αποτελεί μια επέκταση του σημερινού ιστού, η οποία μετατρέπει το μεγάλο απόθεμα πληροφοριών που είναι αναρτημένο στο διαδίκτυο σε γνώση, αποδίδοντας στο διαδικτυακό περιεχόμενο σαφή σημασιολογία [5], [6], [7]. Πρωτοεμφανίστηκε στις αρχές της δεκαετίας του '60 από τον γνωστικό επιστήμονα Allan M. Collins, τον γλωσσολόγο M. Ross Quillian και την ψυχολόγο Elizabeth F. Loftus [1], [2], [3], [4]. Ο όρος, όμως, "σημασιολογικός ιστός" εμνεύστηκε από τον δημιουργό του παγκόσμιου ιστού και διευθυντή του World Wide Web Consortium (W3C), Tim Berners-Lee [8]. Σύμφωνα με αυτόν:

The Semantic Web is an extension of the current Web in which information is given

well-defined meaning, better enabling computers and people to work in cooperation.

. . . a web of data that can be processed directly and indirectly by machines.

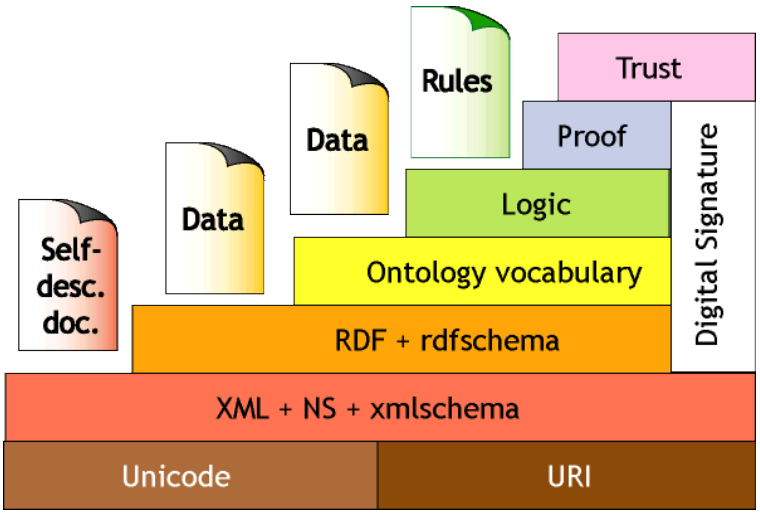
— Tim Berners-Lee, James Hendler, Ora Lassila [8]

Η βασικότερη ιδέα είναι ότι η δημοσιευμένη πληροφορία θα περιέχει μεταδεδομένα (metadata), τα οποία δε θα είναι κατανοητά μόνο από ανθρώπους, αλλά και από μηχανές, δηλαδή η δημιουργία ενός ιστού από δεδομένα ή όπως είναι εκτενέστερα γνωστό “web of data”. Ο απώτερος σκοπός αυτού του εγχειρήματος είναι να προσφέρει περισσότερες δυνατότητες στους υπολογιστές, ώστε να επιτυγχάνουν βέλτιστη συλλογή και επεξεργασία των πληροφοριών. Με αυτόν τον τρόπο είναι δυνατό να αναπτυχθούν συστήματα που θα μπορούν να εγγυηθούν αξιόπιστες ανταλλαγές δεδομένων και πληροφοριών μέσα στο διαδίκτυο.

Συνολικά, οι τεχνολογίες που σχετίζονται με τον σημασιολογικό ιστό, δίνουν τη δυνατότητα στους ανθρώπους να δημιουργούν αποθετήρια δεδομένων (data stores), να δημιουργούν συσχετιζόμενα λεξικά (vocabularies) και να θέτουν τους αντίστοιχους κανόνες για τη διαχείριση αυτών. Στον πυρήνα των τεχνολογιών αυτών ανήκουν εργαλεία όπως τα RDF, SPARQL, OWL, κλπ, τα οποία θα αναλυθούν στις επόμενες ενότητες.

2.1.1 Αρχιτεκτονική

Η αρχιτεκτονική του σημασιολογικού ιστού βασίζεται στη χρήση επιπρόσθετων επιπέδων (layers) για την ενίσχυση του ήδη υπάρχοντος παγκόσμιου ιστού. Αυτά τα επιπλέον επίπεδα αποτελούν τεχνολογικά επίπεδα λειτουργικότητας (functionality), δηλαδή σύνολα γλωσσών προγραμματισμού και τεχνολογιών, τα οποία υποστηρίζονται από τις τεχνολογίες των κατωτέρων επιπέδων [11] και δεν ανταποκρίνονται αυστηρά στην έννοια της αρχιτεκτονικής λογισμικού. Η διαστρωμάτωση των τεχνολογιών στις οποίες βασίζεται ο σημασιολογικός ιστός, αναπαρίσταται γραφικά από τη στοίβα του σημασιολογικού ιστού (semantic web stack), όπως απεικονίζεται στην Εικόνα 1.



Εικόνα 1: Αρχιτεκτονική του σημασιολογικού διαδικτύου κατά τον BernersLee [9].

Στο 1ο επίπεδο αναπαρίσταται η ήδη υπάρχουσα δομή του παγκόσμιου ιστού, πάνω στην οποία θεμελιώνεται ο σημασιολογικός ιστός, δηλαδή: τα καθολικά αναγνωριστικά των πόρων ή URIs (Universal Resource Indicators) για την ονοματολογία και την κωδικοποίηση Unicode (Universal Code) για την καθολική προσπέλαση.

Στο 2ο επίπεδο αναπαρίσταται το περιεχόμενο βάσει του προτύπου XML (eXtensible Markup Language), το οποίο αποτελεί μία μεταγλώσσα σήμανσης, δηλαδή ένα σύνολο κανόνων για το σχεδιασμό μορφών κειμένου που διευκολύνουν τη δόμηση ενός εγγράφου. Η XML δεν επιβάλλει κανέναν σημασιολογικό περιορισμό και προσφέρει τη δυνατότητα δημιουργίας πληροφοριακών πόρων με πολύπλοκη και ευέλικτη δομή.

Στο 3ο επίπεδο πραγματοποιείται η σημασιολογία και περιγραφή των μεταδεδομένων των πληροφοριακών πόρων, ώστε να γίνεται κατανοητή στους ηλεκτρονικούς υπολογιστές. Αυτό επιτυγχάνεται μέσω του πλαισίου περιγραφής πόρων ή RDF (Rersource Description Framework) και του RDF schema, τα οποία θα περιγραφούν αναλυτικότερα στις ενότητες 2.2 και 2.3, αντίστοιχα.

Στο 4ο επίπεδο βρίσκονται οι οντολογίες, οι οποίες θα αναλυθούν στην ενότητα 2.3, και οι διαδικτυακές γλώσσες ορισμού τους.

Στο 5ο επίπεδο βρίσκεται η λογική, η οποία παρέχει τη δυνατότητα εξαγωγής συμπερασμάτων και συλλογισμών βάσει των δομημένων πληροφοριών μία οντότητας.

Στο 6ο επίπεδο αναπαρίσταται η τεκμηρίωση των συμπερασμάτων που έχουν εξαχθεί από δηλώσεις (statements), οι οποίες θα περιλαμβάνουν τεκμήρια εγκυρότητας.

Τέλος, στο 7ο επίπεδο βρίσκεται η εμπιστοσύνη, η οποία έχει σκοπό να εξασφαλίζει την αξιοπιστία των πληροφοριών του σημασιολογικού ιστού που διακινούνται, επεξεργάζονται και συμπεραίνονται, βάσει ψηφιακών υπογραφών και πιστοποιητικών.

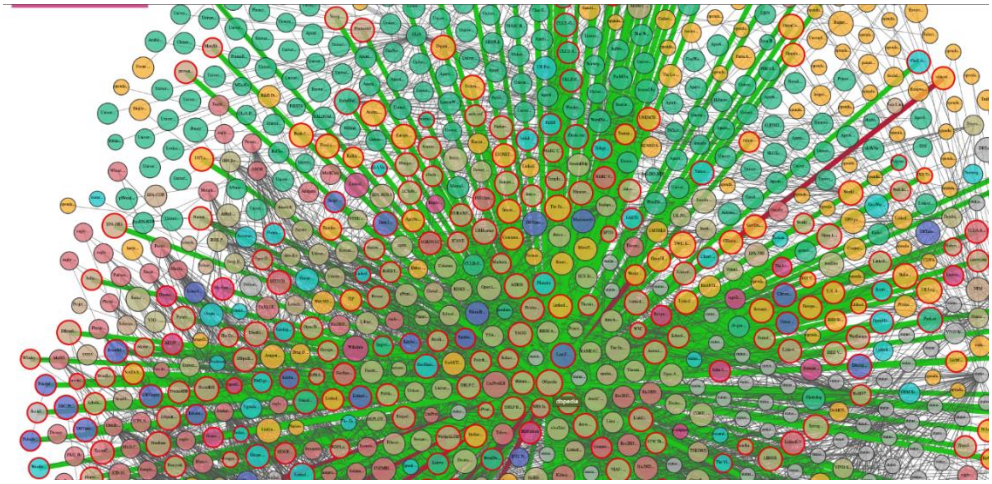
2.1.2 Συνδεδεμένα δεδομένα

Όπως αναφέρθηκε παραπάνω, ο σημασιολογικός ιστός αποτελεί έναν ιστό από δεδομένα, δεδομένα όπως αναγνωριστικός κωδικός ενός προϊόντος, ένας τίτλος βιβλίου, ένα μοναδικό αναγνωριστικό ενός φοιτητή πανεπιστημίου και οποιοδήποτε δεδομένο μπορεί να ενδιαφέρει κάποιο χρήστη του διαδικτύου.

Οι τεχνολογίες του σημασιολογικού ιστού δημιουργούν το κατάλληλο υπόβαθρο πάνω στο οποίο μία εφαρμογή μπορεί να δημιουργήσει ένα ερώτημα (query) για συγκεκριμένα δεδομένα, να οδηγήσει σε κάποιο συμπέρασμα (inference), χρησιμοποιώντας τα κατάλληλα λεξικά.

Η δημιουργία ενός τέτοιου ιστού απαιτεί αφενός μία πληθώρα διαθέσιμων δεδομένων. Αφετέρου, είναι εξίσου σημαντικό τα δεδομένα αυτά να είναι διαθέσιμα με μία συγκεκριμένη και τυποποιημένη μορφή (format) που να τα καθιστά προσβάσιμα και διαχειρίσιμα από τις προαναφερθέντα εργαλεία.

Επιπλέον, θεμελιώδες χαρακτηριστικό του σημασιολογικού ιστού είναι η ύπαρξη σχέσεων (relationships) μεταξύ των δεδομένων και όχι μόνο η απλή αποθήκευσή τους. Αυτή η συλλογή των συσχετιζόμενων δεδομένων αναφέρεται και ως συνδεδεμένα δεδομένα.



Εικόνα 2: Αναπαράσταση των συνδεδεμένων δεδομένων σε μορφή γράφου [16]

Ένα από τα πιο γνωστά σύνολα διασυνδεδεμένων δεδομένων του σημασιολογικού ιστού είναι το dbpedia [16], το οποίο διαθέτει το περιεχόμενο του Wikipedia [15] σε RDF μορφή. Όπως παρουσιάζεται στην Αναπαράσταση των συνδεδεμένων δεδομένων σε μορφή γράφου[16], η πλατφόρμα αυτή πραγματοποιεί πολλές συνδέσεις με άλλους κόμβους του σημασιολογικού ιστού, το οποίο την καθιστά θεμελιώδη δομή. Η μεγάλη σημασία του dbpedia δεν έγκειται μόνο στο γεγονός ότι περικλύει όλα τα δεδομένα του Wikipedia, αλλά και στο ότι δημιουργεί συνδέσεις σε άλλες βάσεις δεδομένων. Με αυτόν τον τρόπο αυτές οι συνδέσεις, οι οποίες παρέχονται σε τριπλέτες RDF, μπορεί να φέρουν στην επιφάνεια εκτενέστερη και πιο ακριβή γνώση κατά τη διαδικασία ανάπτυξης μίας εφαρμογής. Με άλλα λόγια, επιτυγχάνεται ένα είδος διασταύρωσης πληροφορίας που καταλήγει να προσφέρει στο χρήστη μία καλύτερη αλληλεπίδραση.

2.1.3 Μεταδεδομένα

Ένα από τα χαρακτηριστικά του World Wide Web είναι ότι οι πόροι, κατά την προσπέλασή τους, δεν προσφέρουν μόνο δεδομένα, αλλά και την αντίστοιχη επεξήγησή τους, μέσω πληροφορίας που υπάρχει για αυτούς. Η πληροφορία για την πληροφορία είναι γενικά γνωστή ως μεταδεδομένα (metadata). Τα μεταδεδομένα αποτελούν κατανοητή πληροφορία για τις μηχανές σχετικά με τους διαδικτυακού πόρους.

Το Dublin Core [12] αποτελεί ένα από τα δημοφιλέστερα λεξικά μεταδεδομένων. Πιο συγκεκριμένα περιλαμβάνει ένα σύνολο 15 στοιχείων (elements) που εφαρμόζονται σε πόρους πληροφοριών για να την περιγραφή τους. Αυτές οι ιδιότητες περιέχουν πληροφορία όπως «τίτλος», «δημιουργός», «αντικείμενο», «περιγραφή», «εκδότης», «ημερομηνία», «γλώσσα», κλπ.

Το Resource Description Framework (RDF), το οποίο θα αναλυθεί στην ενότητα 2.2, αποτελεί ένα μοντέλο δεδομένων για μεταδεδομένα το οποίο σχεδιάστηκε από World Wide Web Consortium (W3C).

2.2 Πλαίσιο περιγραφής πόρων - RDF

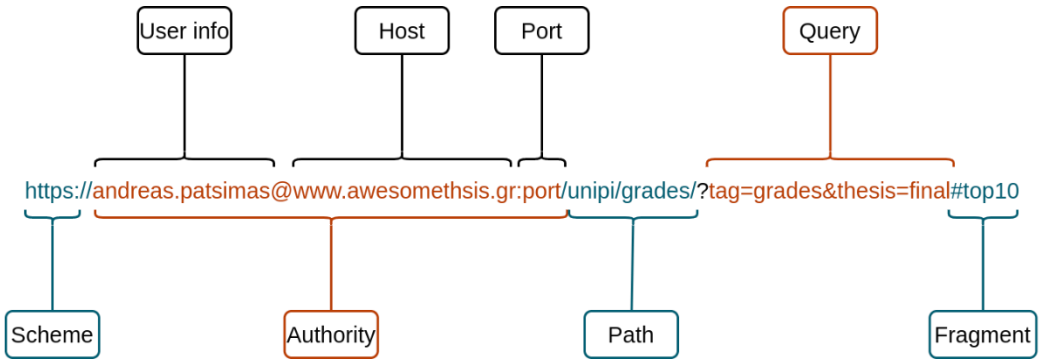
Το πλαίσιο περιγραφής πόρων – RDF (Resource Description Framework) [20] αποτελεί αδιάσπαστο συστατικό στοιχείο του σημασιολογικού ιστού, καθώς καθορίζει τη μορφή των δεδομένων. Μέσω του RDF, γίνεται εφικτή η μετατροπή μεταξύ δύο διαφορετικών τύπων και επιτρέπεται η άμεση πρόσβαση σε υπάρχουσες βάσεις δεδομένων (σχεσιακές, XML, HTML, κλπ).

Για την πρόσβαση στα δεδομένα αυτά και την τυποποίηση της ενέργειας αυτής, είναι σημαντικός ο καθορισμός τερματικών σημείων (endpoints), τα οποία περιμένουν ως είσοδο τα αντίστοιχα ερωτήματα (queries). Με αυτόν τον τρόπο, η πρόσβαση στις πηγές δεδομένων πραγματοποιείται μέσω μίας καλύτερα δομημένης και κατάλληλης διαδικασίας.

2.2.1 Ομοιόμορφο αναγνωριστικό πόρων – URI

Ένα ομοιόμορφο αναγνωριστικό πόρων – URI (Uniform Resource Identifier) αποτελεί μία μοναδική ακολουθία χαρακτήρων, η οποία ταυτοποιεί έναν λογικό ή φυσικό πόρο που χρησιμοποιείται από τεχνολογίες ιστού. Τα URIs μπορούν να χρησιμοποιηθούν για να ταυτοποιήσουν το οτιδήποτε, συμπεριλαμβανομένου αντικείμενα του πραγματικού κόσμου, όπως ανθρώπους ή τοποθεσίες, έννοιες, ή πόρους πληροφοριών όπως ιστοσελίδες ή βιβλία. Κάποια URIs προσφέρουν έναν τρόπο εντοπισμού και ανάκτησης πληροφοριακών πόρων σε ένα δίκτυο (είτε το διαδίκτυο ή σε ένα άλλο ιδιωτικό δίκτυο, όπως σε ένα σύστημα αρχείων ενός υπολογιστή ή σε Intranet)· αυτά είναι ομοιόμορφοι εντοπιστές πόρων – URLs (Uniform Resource Locators). Ένα URL προσφέρει την τοποθεσία ενός πόρου. Ένα URI ταυτοποιεί έναν πόρο μέσω του ονόματός του σε μία

συγκεκριμένη τοποθεσία ή URL. Άλλα URIs προσφέρουν μόνο ένα μοναδικό όνομα, χωρίς έναν τρόπο εντοπισμού ή ανάκτησης του πόρου ή της πληροφορίας σχετικά με αυτό· Αυτά ονομάζονται ομοιόμορφα ονόματα πόρων - URNs (Uniform Resource Names). Οι τεχνολογίες ιστού που χρησιμοποιούν URIs δεν περιορίζονται σε προγράμματα περιήγησης ιστού (web browsers). Τα URIs χρησιμοποιούνται για να ταυτοποιήσουν οτιδήποτε περιγράφεται μέσω του RDF. Για παράδειγμα έννοιες που είναι μέρος μία οντολογίας και ορίζονται μέσω της γλώσσας οντολογίας διαδικτύου – OWL, και ανθρώπους που περιγράφονται χρησιμοποιώντας το λεξικό Friend of a Friend, θα είχαν κάθε ένα από αυτά ένα μοναδικό URI.



Εικόνα 3: Παράδειγμα γενική μορφής URI.

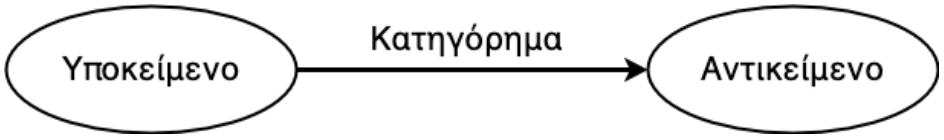
Στην Εικόνα 3 παρουσιάζονται τα συστατικά μέρη ενός URI γενικής μορφής. Σε περίπτωση που η αναζήτηση πληροφοριών γίνεται σε πηγή με περιορισμένη πρόσβαση, το τμήμα που σημειώνεται ως «Authority» είναι απαραίτητο, καθώς περιλαμβάνει όχι μόνο τη διεύθυνσή της, αλλά και τα απαραίτητα στοιχεία του χρήστη. Το τμήμα «Path» χρησιμοποιείται όταν η επιθυμητή πληροφορία δεν βρίσκεται στην αρχική διαδρομή της σχετικής διεύθυνσης. Όταν η αναζήτηση απαιτεί πιο στοχευμένα αποτελέσματα, τότε γίνεται χρήση του αντίστοιχου επερωτήματος (query).

2.2.2 Τριπλέτες

Το RDF βασίζεται στην αρχή ότι τρία κομμάτια πληροφορίας είναι αρκετά προκειμένου να οριστεί πλήρως ένα ξεχωριστό bit γνώσης. Εντός του προσδιορισμού του RDF, μία RDF τριπλέτα (triple) [19] τεκμηριώνει αυτά τα τρία κομμάτια πληροφορίας με έναν συνεπή τρόπο, ο οποίος ιδανικά επιτρέπει ταυτόχρονα και σε ανθρώπους και σε μηχανές να λαμβάνουν τα ίδια δεδομένα.

Από τα τρία κομμάτια πληροφοριών, το πρώτο είναι το υποκείμενο (subject). Μία ιδιότητα όπως όνομα μπορεί να ανήκει σε ένα σκύλο, γάτα, βιβλίο, φυτό, άτομο, αυτοκίνητο, έθνος ή έντομο. Προκειμένου να γίνει πεπερασμένο ένα τόσο απέραντο σύμπαν, πρέπει να τεθούν όρια, και αυτό επιτυγχάνει στο υποκείμενο για το RDF. Το δεύτερο κομμάτι πληροφορίας είναι ο τύπος ιδιότητας (property type) ή απλώς απλή ιδιότητα (property). Υπάρχουν διάφορα χαρακτηριστικά για οποιοδήποτε μεμονωμένο υποκείμενο, Για παράδειγμα, έχω ένα φύλο, ύψος, χρώμα μαλλιού, χρώμα ματιών, πτυχίο πανεπιστημίου, σχέση, κλπ. Προκειμένου να προσδιοριστεί για ποιο χαρακτηριστικό του υποκειμένου υπάρχει ενδιαφέρον, πρέπει να γίνει εστίαση σε μία ιδιότητα. Η τομή του υποκειμένου με την ιδιότητα οδηγεί στο τρίτο κομμάτι της πληροφορίας, την τιμή (value). Δηλαδή «Εγώ (υποκείμενο) έχω ένα όνομα (ιδιότητα), το οποίο είναι Αντρέας (τιμή ιδιότητας)».

Στο RDF, όπως απεικονίζεται στην Εικόνα 4 το υποκείμενο είναι αυτό που περιγράφεται σε όρους RDF, δηλαδή ένας πόρος που ταυτοποιείται από ένα URI. Το κατηγόρημα (predicate) είναι ο τύπος ιδιότητας του πόρου, όπως ένα χαρακτηριστικό, μία σχέση ή μία ιδιότητα. Επιπρόσθετα του υποκειμένου και κατηγορουμένου, η προδιαγραφή εισάγει και ένα τρίτο συστατικό, το αντικείμενο (object). Στο RDF, το αντικείμενο παίρνει την τιμή (μπορεί να είναι λεκτικό, αριθμητική τιμή ή πόρος) της ιδιότητας του πόρου (υποκείμενο) του συγκεκριμένου υποκειμένου.



Εικόνα 4: Το βασικό μοντέλο σημασιολογικής τριπλέτας.

2.2.3 SPARQL

Όταν αναφερόμαστε σε ένα ερώτημα (query) για το σημασιολογικό ιστό, εννοούμε όλες τις τεχνολογίες και τα πρωτόκολλα που συνδυάζονται για ανάκτηση πληροφοριών από τον ιστό των δεδομένων. Όπως ήδη έχει αναφερθεί, το RDF αποτελεί τη βάση για τη δομή και τη διασύνδεση των δεδομένων. Πληθώρα τεχνολογιών επιτρέπουν την ενσωμάτωση των δεδομένων σε έγγραφα ή την έκθεση των δεδομένων που υπάρχουν σε βάση δεδομένων ή διάθεση αυτών σε RDF αρχεία.

Ο σημασιολογικός ιστός, όπως οι άλλες βάσεις δεδομένων (σχεσιακές, XML, κλπ), απαιτεί και το αντίστοιχο προσαρμοσμένο συντακτικό για τα ερωτήματα που αφορούν δεδομένα μορφής RDF. Τέτοια προσαρμοσμένα ερωτήματα παρέχονται από τη γλώσσα επερωτημάτων SPARQL (Simple Protocol and RDF Query Language) [17], η οποία χρησιμοποιεί τα ανάλογα πρωτόκολλα.

Πρακτικά μιλώντας, τα ερωτήματα σε SPARQL βασίζονται σε μοτίβα τριπλετών. Το RDF μπορεί να θεωρηθεί ένα σύνολο σχέσεων μεταξύ των διάφορων πόρων δεδομένων. Αυτά τα ερωτήματα SPARQL παρέχουν ένα ή περισσότερα μοτίβα, σχετικά με τις σχέσεις αυτές. Η μηχανή SPARQL επιστρέφει τις πηγές για όλες τις τριπλέτες που ταιριάζουν σε αυτά τα μοτίβα.

Οι χρήστες του σημασιολογικού ιστού μπορούν να εξάγουν σύνθετες πληροφορίες χρησιμοποιώντας την SPARQL. Οι πληροφορίες αυτές επιστρέφονται σε μορφή πίνακα, ο οποίος μπορεί να ενσωματωθεί και σε άλλες ιστοσελίδες. Με αυτόν τον τρόπο, η SPARQL παρέχει ένα ισχυρό εργαλείο για τη δημιουργία παραδείγματος χάριν μίας μηχανής αναζήτησης που περιλαμβάνει δεδομένα που πηγάζουν από το σημασιολογικό ιστό.

Για παράδειγμα, ας υποθέσουμε την τριπλέτα «Batman isA superhero». Χρησιμοποιώντας την SPARQL, μπορούμε να συντάξουμε ένα ερώτημα ως εξής:

Batman isA ?attribute,

όπου το ?attribute εκφράζει μία μεταβλητή.

Η μηχανή της SPARQL αναζητεί μέσα στα διαθέσιμα δεδομένα και επιστρέφει την τιμή «superhero» ως μία πιθανή τιμή του γνωρίσματος (attribute), το οποίο κατ' επέκταση συνιστά μία πιθανή απάντηση στο ερώτημα. Το σύνολο των διαθέσιμων δεδομένων ίσως να περιλαμβάνει επίσης την τριπλέτα «Batman isA human». Σε αυτή την περίπτωση η τιμή «human» αποτελεί επίσης μία πιθανή τιμή του γνωρίσματος και επομένως θα πρέπει να επιστραφεί και αυτό ως απάντηση στο ερώτημα. Παρέχοντας τέτοιες πολλαπλές τριπλέτες μπορούν να συνταχθούν σύνθετα ερωτήματα, τα οποία μπορούν να χρησιμοποιηθούν από μία εφαρμογή.

2.3 RDFS, λεξικά και οντολογίες

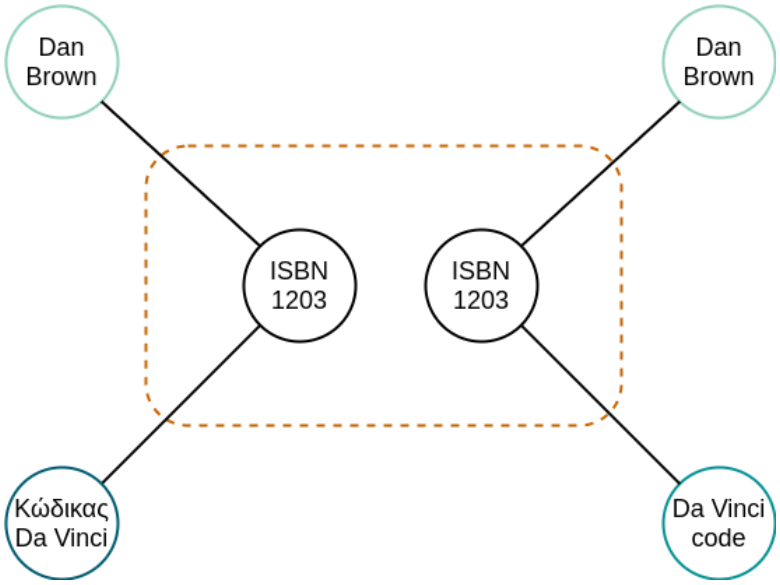
Όπως προαναφέρθηκε, το RDF μοντέλο δεδομένων ορίζει ένα απλό μοντέλο για την περιγραφή των αλληλεπιδράσεων μεταξύ των πόρων, όσον αφορά τις ονομαστικές ιδιότητες και τιμές. Οι RDF ιδιότητες μπορούν να θεωρηθούν ως ιδιότητες των πόρων

και με αυτή την έννοια αντιστοιχούν σε παραδοσιακά ζεύγη χαρακτηριστικών-τιμών. Επιπλέον, οι ιδιότητες RDF εκπροσωπούν τις σχέσεις μεταξύ των πόρων. Επομένως, το RDF μοντέλο δεδομένων μπορεί να θυμίζει ένα διάγραμμα οντότητα-σχέσεων. Όμως, το RDF μοντέλο δεδομένων δεν παρέχει κάποιο μηχανισμό για τον ορισμό των σχέσεων μεταξύ αυτών των ιδιοτήτων και των άλλων πόρων. Αυτός είναι ο ρόλος του RDF Schema ή RDFS [13]. Επομένως, το RDFS μπορεί να θεωρηθεί ως μία περιγραφική γλώσσα RDF λεξιλογίων, καθώς χρησιμοποιείται για να δημιουργήσει λεξικά, τα οποία εξυπηρετούν στο να είναι οι διανεμημένες πληροφορίες πιο φιλικές και επεξεργάσιμες προς τις μηχανές.

Τα λεξικά στο πλαίσιο του σημασιολογικού ιστού χρησιμοποιούνται για τον ορισμό εννοιών και συσχετίσεων που περιγράφουν και αντιπροσωπεύουν τις περιοχές ενδιαφέροντος. Βασική τους χρήση είναι η ταξινόμηση των όρων που μπορούν να χρησιμοποιηθούν σε μία συγκεκριμένη εφαρμογή, να χαρακτηρίσουν πιθανές συσχετίσεις και να καθορίσουν πιθανούς περιορισμούς κατά τη χρήση αυτών των όρων. Πρακτικά τα λεξικά μπορεί να είναι πολύ σύνθετα, περιλαμβάνοντας χιλιάδες όρους, ή πολύ απλά διαθέτοντας μόνο έναν ή δύο.

Είναι σύνηθες τα λεξικά να αναφέρονται και σαν οντολογίες, καθώς ο διαχωρισμός τους δεν είναι πάντα ξεκάθαρος. Είθισται η λέξη οντολογία να χρησιμοποιείται για πιο σύνθετους και πιθανούς επίσημους όρους, ενώ τα λεξικά χρησιμοποιούνται όταν ένας τέτοιος φορμαλισμός δεν είναι απαραίτητος.

Η χρήση των οντολογιών [20] στο σημασιολογικό ιστό εξυπηρετεί στη σύνθεση δεδομένων, σε περιπτώσεις όπου υπάρχουν ανακρίβειες σε συγκεκριμένους όρους που χρησιμοποιούνται σε διαφορετικά σύνολα δεδομένων, ή όταν πρόσθετες πληροφορίες μπορούν να οδηγήσουν στη δημιουργία νέων σχέσεων ή και αλλού. Για παράδειγμα, ας υποθέσουμε ότι έχουμε μία οντότητα που περιγράφει ένα συγκεκριμένο βιβλίο. Η οντότητα αυτή συνδέεται με διάφορες οντότητες που σχετίζονται με την περιγραφή του βιβλίου. Αναφορικά με την πρώτη περίπτωση (ανακρίβεια δεδομένων), ο συγγραφέας μπορεί να αναφέρεται είτε ως «δημιουργός» είτε ως «συγγραφέας» ,δημιουργώντας έτσι ασυνέπεια Όσον αφορά τη δεύτερη περίπτωση (νέα γνωρίσματα ή σχέσεις), μπορεί να υπάρχουν πολλαπλές μεταφράσεις του βιβλίου αυτού, με τον ίδιο αναγνωριστικό κωδικό (ISBN). Η συνένωση αυτών των οντοτήτων με κοινό ISBN μπορεί να προσφέρει ευκολία στην πρόσβαση και επέκταση της πληροφορίας που σχετίζεται με το βιβλίο. Ειδικότερα, με την πρόσβαση στην οντότητα μέσω του μοναδικού ISBN του βιβλίου, ο χρήστης θα έχει τη δυνατότητα να έχει πρόσβαση στους διαφορετικούς του τίτλους, ανάλογα με τη χώρα.



Εικόνα 5: Παράδειγμα συνένωσης διαφορετικών στιγμιότυπων της ίδιας οντότητας.

Χαρακτηριστικό παράδειγμα είναι η χρήση των οντολογιών για την οργάνωση της γνώσης. Βιβλιοθήκες, μουσεία, εφημερίδες, εταιρείες, μέσα κοινωνικής δικτύωσης και άλλες κοινότητες που διαχειρίζονται εκτενή όγκο δεδομένων μπορούν πλέον να χρησιμοποιήσουν τις οντολογίες ώστε να αξιοποιήσουν τις απεριόριστες δυνατότητες των διασυνδεδεμένων δεδομένων.

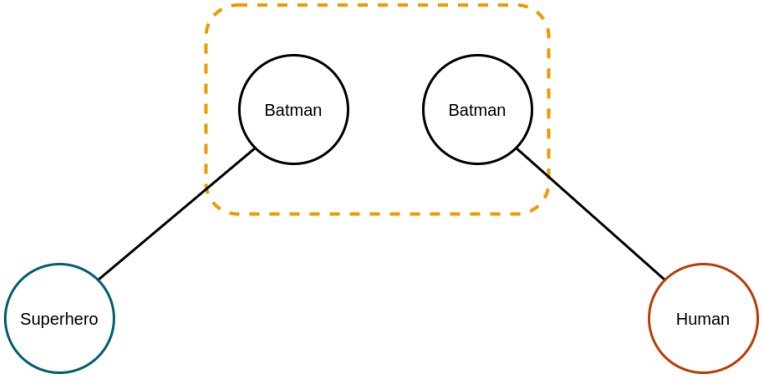
2.3.1 Συμπεράσματα

Τα συμπεράσματα στο σημασιολογικό ιστό μπορούν να χαρακτηριστούν με την ανακάλυψη νέων σχέσεων. Τα δεδομένα στο σημασιολογικό ιστό μοντελοποιούνται ως ένα σύνολο σχέσεων ανάμεσα σε πόρους δεδομένων. Με τον όρο συμπέρασμα περιγράφονται οι αυτοματοποιημένες διαδικασίες με τις οποίες δημιουργούνται νέες σχέσεις με βάση τα δεδομένα που υπάρχουν και πρόσθετες πληροφορίες που βρίσκονται στις υπάρχουσες οντολογίες. Αυτές οι νέες σχέσεις είτε προστίθενται στο σύνολο των δεδομένων είτε επιστρέφονται κατά την επιστροφή της απάντησης ενός ερωτήματος (query).

Οι πηγές των πρόσθετων πληροφοριών ορίζονται από ένα σύνολο οντολογιών και ένα σύνολο κανόνων. Και τα δύο αυτά εργαλεία στοχεύουν στο σχηματισμό κατάλληλων τεχνικών παρουσίασης αυτής της πληροφορίας. Οι οντολογίες περιλαμβάνουν μεθόδους ταξινόμησης δίνοντας έμφαση στον ορισμό κλάσεων και υπο-κλάσεων στις οποίες εντάσσονται οι ανεξάρτητες πηγές και στις αλληλεπιδράσεις μεταξύ τους. Αντίθετα, το

σύνολο των κανόνων αποσκοπεί στον ορισμό κατάλληλων μηχανισμών για την ανακάλυψη και δημιουργία νέων σχέσεων με βάση τις ήδη υπάρχουσες.

Για την καλύτερη κατανόηση της ιδιότητας συμπεράσματος του σημασιολογικού ιστού, θα χρησιμοποιηθεί το αντίστοιχο παράδειγμα της ενότητας 2.2.3. Δεδομένης της τριπλέτας «Batman isA superhero», μία οντολογία μπορεί να δηλώσει ότι «Batman is also a human». Πρακτικά αυτό σημαίνει ότι ένα πρόγραμμα του σημασιολογικού ιστού που αντιλαμβάνεται την έννοια «X is also Y» μπορεί να προσθέσει τη δήλωση «Batman isA human» στο σύνολο των σχέσεων παρ' όλο που δεν υπήρχε στα αρχικά δεδομένα. Με αυτή την έννοια μπορεί να θεωρηθεί ότι αυτή η νέα σχέση «ανακαλύφθηκε».



Εικόνα 6: Παράδειγμα «ανακάλυψης» μίας νέας σχέσης.

2.3.2 Γλώσσα οντολογίας διαδικτύου (OWL)

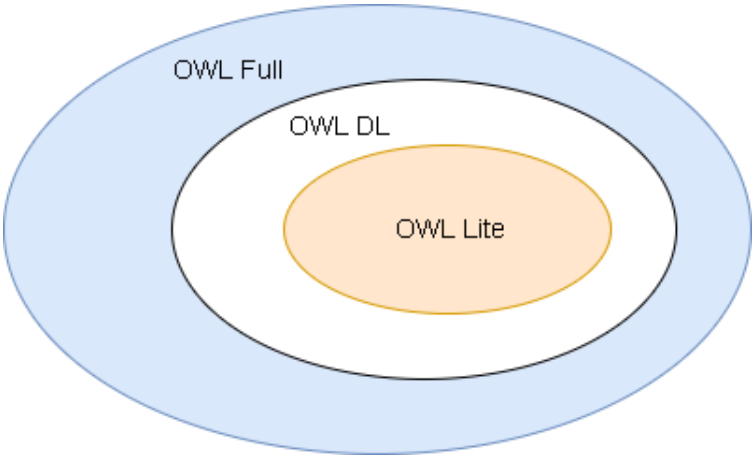
Η γλώσσα οντολογίας διαδικτύου (OWL) [18] αποτελεί μία γλώσσα σημασιολογικού ιστού, η οποία σχεδιάστηκε από τον W3C με σκοπό να αναπαριστά πλούσια και περίπλοκη γνώση για πράγματα, ομάδες πραγμάτων και σχέσεις μεταξύ πραγμάτων. Η OWL είναι μία γλώσσα που βασίζεται σε υπολογιστική λογική, έτσι ώστε η γνώση που εκφράζεται σε OWL να μπορεί να είναι εκμεταλλεύσιμη από προγράμματα υπολογιστών, για παράδειγμα για την επαλήθευση της εγκυρότητας αυτής της γνώσης ή για να γίνει η σαφής γνώση άρρητη. Τα έγγραφα OWL, γνωστά ως οντολογίες, μπορούν να δημοσιευτούν στον παγκόσμιο ιστό και μπορούν να αναφέρουν ή να αναφέρονται σε άλλες OWL οντολογίες.

Όπως παρουσιάζεται στην Εικόνα 7 η OWL παρέχει τρεις πιο εκφραστικές υπογλώσσες, οι οποίες σχεδιάστηκαν για χρήση από συγκεκριμένες κοινότητες προγραμματιστών και χρηστών.

● Η OWL Lite υποστηρίζει αυτούς του χρήστες που πρωτίστως χρειάζονται μία ιεραρχία ταξινόμησης και απλούς περιορισμούς. Για παράδειγμα, καθώς υποστηρίζει περιορισμούς πληθάριθμου (cardinality), επιτρέπει μόνο τιμές πληθάριθμου 0 και 1. Η παροχή εργαλείων υποστήριξης είναι απλούστερη για την OWL Lite από ότι είναι για τις πιο ακριβές συγγενείς της, και η OWL Lite παρέχει μια γρήγορη μετακίνηση μονοπατιών για δομές λεξικών και άλλες ταξινομίες.

● Η OWL DL υποστηρίζει αυτούς τους χρήστες που θέλουν τη μέγιστη εκφραστικότητα, ενώ παράλληλα διατηρούν υπολογιστική πληρότητα (όλα τα αποτελέσματα είναι εγγυημένα υπολογίσιμα) και αποφασιστικότητα (όλοι οι υπολογισμοί θα ολοκληρωθούν σε πεπερασμένο χρόνο). Η OWL DL περιλαμβάνει όλες τις γλωσσικές κατασκευές της OWL, αλλά μπορούν να χρησιμοποιηθούν μόνο υπό συγκεκριμένους περιορισμούς, όπως για παράδειγμα, ενώ μία κλάση μπορεί να είναι υποκλάση πολλών κλάσεων, μία κλάση δεν μπορεί να είναι στιγμιότυπο μία άλλης κλάσης. Η OWL DL απέκτησε το όνομά της λόγω της αντιστοιχίας της με (Description Logics), ένα πεδίο έρευνας που έχει μελετήσει τις λογικές που αποτελούν την επίσημη βάση της OWL.

● Η OWL Full προορίζεται για χρήστες που θέλουν μέγιστη εκφραστικότητα και συντακτική ελευθερία του RDF χωρίς υπολογιστικές εγγυήσεις. Για παράδειγμα, στην OWL Full μία κλάση μπορεί να αντιμετωπιστεί ταυτόχρονα ως μία συλλογή ατόμων και ως ένα άτομο από μόνο του. Η OWL Full επιτρέπει σε μία οντολογία να αυξήσει το νόημα ενός προ-ορισμένου (RDF ή OWL) λεξιλογίου. Είναι απίθανο οποιοδήποτε λογισμικό συλλογιστικής να καταφέρει να υποστηρίζει πλήρη συλλογιστική για κάθε χαρακτηριστικό της OWL Full.



Εικόνα 7: Οι τρεις υπογλώσσες της γλώσσας οντολογίας διαδικτύου (OWL).

3 Περιγραφή υλοποίησης

3.1 Απαιτήσεις και στόχοι

Βασικός στόχος της παρούσας εργασίας είναι η συλλογή των δεδομένων που σχετίζονται με Covid-19, το περιβάλλον, την μετακίνηση των ανθρώπων, κ.α και η κατασκευή μίας δομημένης οντολογίας που βασίζεται στις αρχές του σημασιολογικού ιστού. Επόμενο βήμα αποτελεί η δημιουργία των κατάλληλων SPARQL queries ώστε να απαντηθούν ερωτήματα που μπορούν να βοηθήσουν στην κατανόηση της πορείας της πανδημίας.

3.1.1 Ερωτήματα σε φυσική γλώσσα

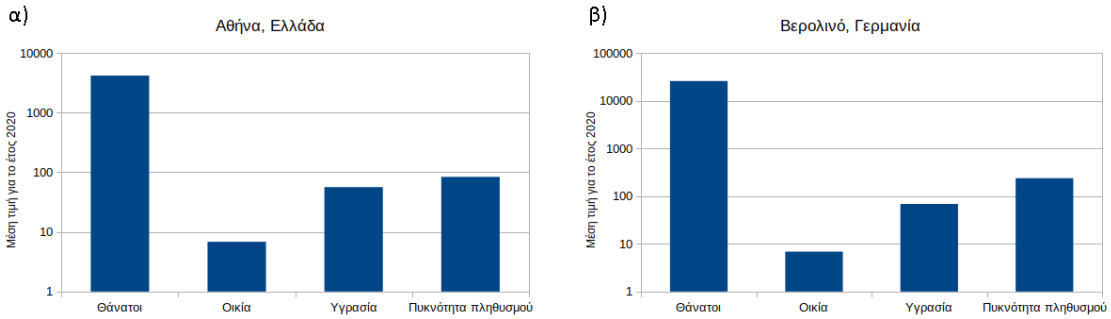
Όπως αναφέρθηκε προηγουμένως, προκειμένου να δοθούν απαντήσεις σχετικά με την πανδημία του Covid-19 και πώς σχετίζεται με τα υπόλοιπα δεδομένα, είναι απαραίτητο να τεθούν και τα κατάλληλα ερωτήματα – queries. Για το λόγο αυτό, δημιουργήθηκαν ενδεικτικά τα παρακάτω ερωτήματα σε φυσική γλώσσα, με σκοπό την μετέπειτα προσπάθεια απάντησή τους:

1) Πώς επηρεάστηκε ο αριθμός των νέων ημερήσιων κρουσμάτων σε κάθε χώρα ανάλογα με την αυστηρότητα των μέτρων.
2) Πώς επηρεάστηκε ο αριθμός των νέων ημερήσιων κρουσμάτων σε κάθε χώρα ανάλογα με την αυστηρότητα των μέτρων και όταν η υγρασία της ατμόσφαιρας ήταν υψηλή.
3) Πώς επηρεάστηκαν οι μετακινήσεις προς τους χώρους εργασίας σε κάθε χώρα ανάλογα με την αυστηρότητα των μέτρων.
4) Πώς επηρεάστηκαν οι μετακινήσεις προς τους χώρους εργασίας σε κάθε χώρα όταν η υγρασία της ατμόσφαιρας ήταν υψηλή.
5) Πώς επηρεάστηκαν οι μετακινήσεις προς τα πάρκα/γήπεδα σε κάθε χώρα ανάλογα με την αυστηρότητα των μέτρων.
6) Πώς επηρεάστηκαν οι μετακινήσεις προς τα πάρκα/γήπεδα κατά τις περιόδους που υγρασία της ατμόσφαιρας ήταν υψηλή.
7) Πώς επηρεάστηκε ο αριθμός των νέων ημερήσιων κρουσμάτων σε κάθε χώρα ανάλογα με τον αριθμό νοσοκομείων της.
8) Πώς επηρεάστηκε ο αριθμός των νέων ημερήσιων κρουσμάτων σε κάθε χώρα ανάλογα με τον αριθμό νοσοκομείων της και την πυκνότητα του πληθυσμού της.

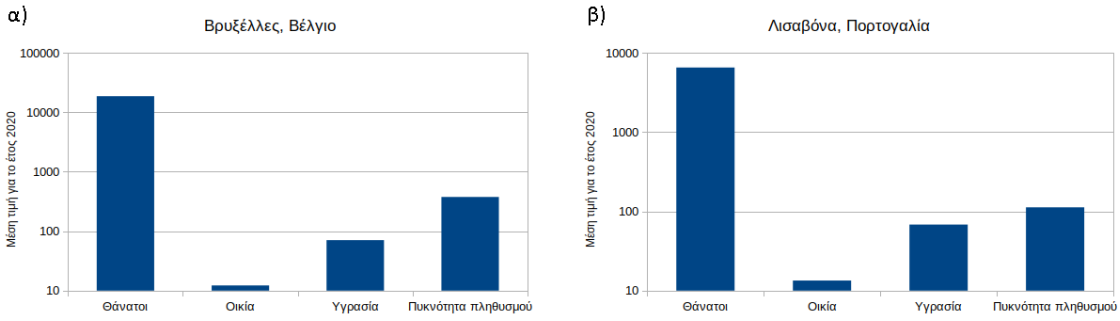
3.2 Συλλογή δεδομένων

Η συλλογή δεδομένων βασίστηκε σε δύο κατηγορίες πηγών, η πρώτη σχετίζεται περισσότερο με εξειδικευμένες πηγές δεδομένων και η δεύτερη με δεδομένα που προέρχονται από τον σημασιολογικό ιστό. Τα δεδομένα που συγκεντρώθηκαν από την πρώτη κατηγορία αφορούν το έτος 2020 και επικεντρώνονται στην πορεία της πανδημίας του ιού Covid-19, στην ατμοσφαιρική ρύπανση, καθώς και σε βασικές μετακινήσεις των ανθρώπων.

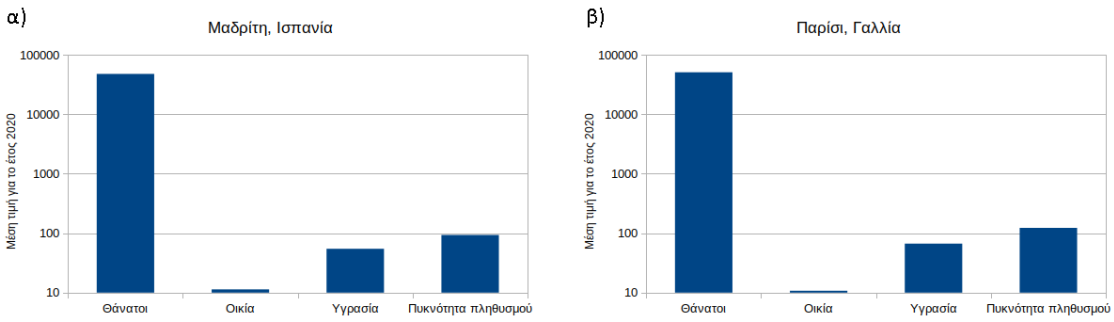
Στις Εικόνες 8 - 11 παρουσιάζεται η μέση τιμή των συλλεχθέντων δεδομένων για το έτος 2020 σχετικά με τους ημερήσιους θανάτους, την παραμονή στο σπίτι, την υγρασία της ατμόσφαιρας και την πυκνότητα πληθυσμού, για 8 διαφορετικές ευρωπαϊκές πόλεις.



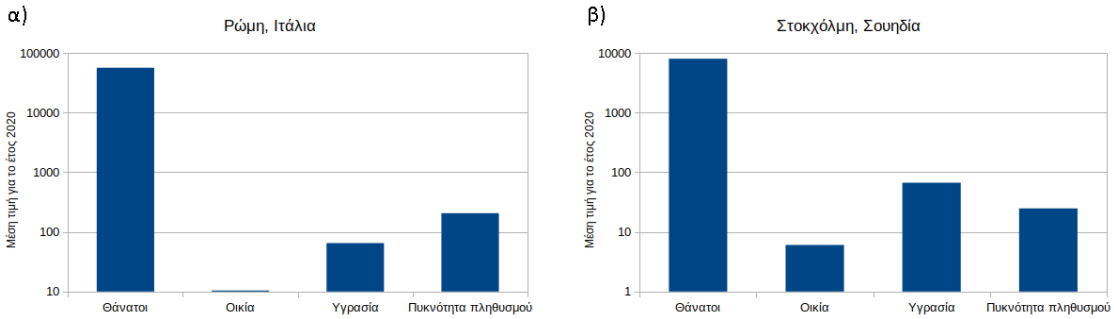
Εικόνα 8: Μέση τιμή ορισμένων δεδομένων για τις πόλεις της Αθήνας (α) και του Βερολίνου (β).



Εικόνα 9: Μέση τιμή ορισμένων δεδομένων για τις πόλεις των Βρυξελλών (α) και της Λισαβόνας (β).



Εικόνα 10: Μέση τιμή ορισμένων δεδομένων για τις πόλεις της Μαδρίτης (α) και του Παρισιού (β).



Εικόνα 11: Μέση τιμή ορισμένων δεδομένων για τις πόλεις της Ρώμης (α) και της Στοκχόλμης (β).

3.2.1 Covid-19

Τα δεδομένα σχετικά με Covid-19 που χρησιμοποιήθηκαν στην παρούσα εργασία συλλέχθηκαν και δημοσιεύτηκαν από το νοσοκομείο Johns Hopkins [21], ενώ η διαχείριση και συντήρησή τους πραγματοποιείται από τον οργανισμό ourworldindata [22]. Εκτός από τον ημερήσιο αριθμό κρουσμάτων και θανάτων, το ίδιο σύνολο περιέχει και δεδομένα σχετικά με τη νοσηλεία ασθενών, με τον αριθμό ανθρώπων που έχουν υποβληθεί σε διαγνωστικό έλεγχο, με τον αριθμό εμβολιασμών, καθώς και με διάφορους δείκτες που αφορούν για παράδειγμα καρδιαγγειακές παθήσεις, και άλλα. Η πλατφόρμα του οργανισμού ourworldindata παρέχει μία σειρά από χρήσιμα εργαλεία απεικόνισης της συσχετιζόμενης πληροφορίας.

Πίνακας 1: Δεδομένα που σχετίζονται με τη μόλυνση του αέρα.

Covid-19	New cases
	New deaths

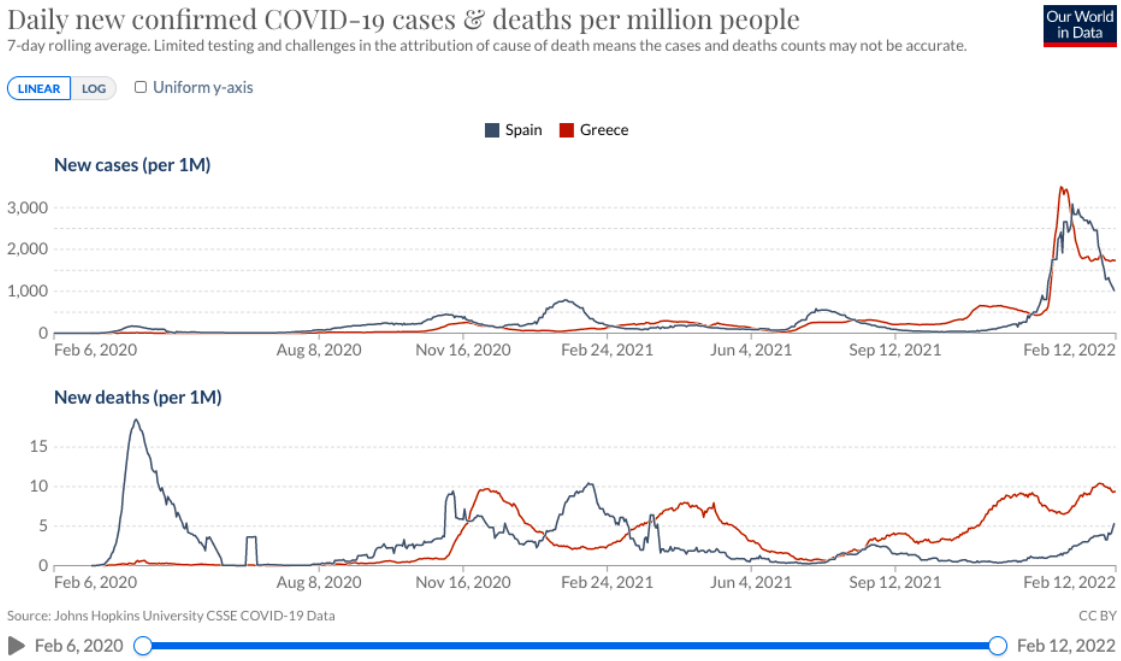
Στην Εικόνα 12 παρουσιάζεται ένας πίνακας σχετικά με τα ημερήσια κρούσματα και ημερήσιους θανάτους ανά εκατομμύριο ανθρώπων. Έτσι ο χρήστης μπορεί να παρατηρήσει για παράδειγμα τις τιμές των νέων κρουσμάτων στην Ελλάδα σε δύο διαφορετικές ημερομηνίες, καθώς επίσης και την απόλυτη και σχετική διαφορά μεταξύ τους.

Μία εναλλακτική και πιο αναλυτική απεικόνιση της ίδια πληροφορίας, παρουσιάζεται στην Εικόνα 13. Πιο συγκεκριμένα, το εργαλείο της εικόνας αυτής δίνει τη δυνατότητα στο χρήστη να συγκρίνει την πορεία της πανδημίας με βάση των αριθμών νέων κρουσμάτων και θανάτων για πολλές ξεχωριστές χώρες. Επίσης, χρησιμοποιώντας αυτό το εργαλείο ο χρήστης μπορεί να μελετήσει διαφορετικά δεδομένα του συνόλου που αφορούν την ίδια χώρα. Αντίστοιχα στην Εικόνα 14 παρουσιάζεται μία σύγκριση του ημερήσιου αριθμού κρουσμάτων μεταξύ πολλαπλών χωρών. Τέλος στην Εικόνα 15 απεικονίζεται ο ημερήσιος αριθμός κρουσμάτων σε παγκόσμια κλίμακα για μία συγκεκριμένη ημέρα επιλεγμένη από το χρήστη σε μορφή heatmap. Ο αριθμός κρουσμάτων απεικονίζεται μέσω της έντασης του χρώματος και το σκουρότερο υποδηλώνει μεγαλύτερο αριθμό, ενώ το πιο ανοιχτόχρωμο, μικρότερο.

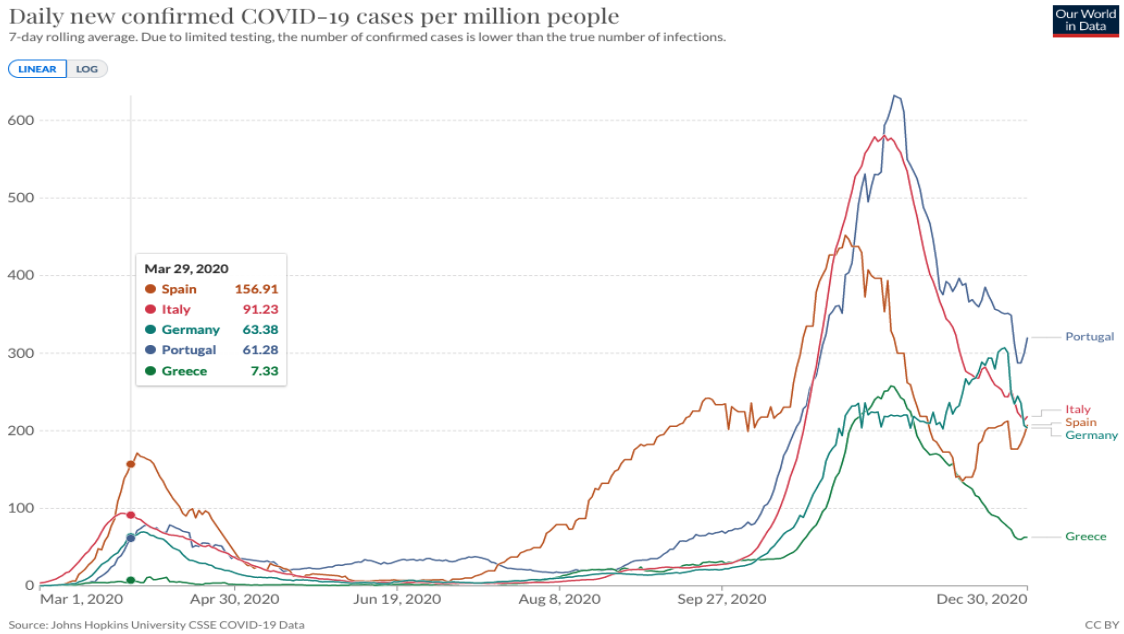
	New cases (per 1M)				Nev	
Country	Jan 28, 2020	Feb 12, 2022	Absolute Change	Relative Change	Jan 28, 2020	Feb 12, 2022
Gabon	Mar 19, 2020 0.06	11.22	+11.16	+17,711%	Mar 20, 2020 0.06	0.00
Gambia	Mar 22, 2020 0.06	1.78	+1.72	+3,025%	Mar 23, 2020 0.06	0.06
Georgia	Mar 2, 2020 0.11	5,127.64	+5,127.53	+4,747,718%	Apr 4, 2020 0.04	12.28
Germany	Feb 1, 2020 0.01	2,258.32	+2,258.30	+16,130,736%	Mar 9, 2020 <0.01	2.07
Ghana	Mar 19, 2020 0.05	Feb 7, 2022 3.74	+3.69	+7,382%	Mar 21, 2020 <0.01	0.03
Gibraltar	Mar 9, 2020 4.24	Feb 11, 2022 3,464.26	+3,460.02	+81,604%	Nov 11, 2020 4.24	0.00
Greece	Mar 2, 2020 0.10	1,740.76	+1,740.67	+1,813,194%	Mar 11, 2020 0.01	9.37
Greenland	Mar 21, 2020 5.02	964.64	+959.62	+19,101%	Dec 27, 2021 2.51	2.51
Grenada	Mar 27, 2020 8.85	515.73	+506.89	+5,729%	Jan 3, 2021 1.26	2.53
Guatemala	Mar 19, 2020 0.07	188.26	+188.19	+268,843%	Mar 19, 2020 <0.01	1.03
[illegible]	Mar 18, 2020	[illegible]	[illegible]	[illegible]	Apr 15, 2020	[illegible]

Jan 28, 2020 — Feb 12, 2022

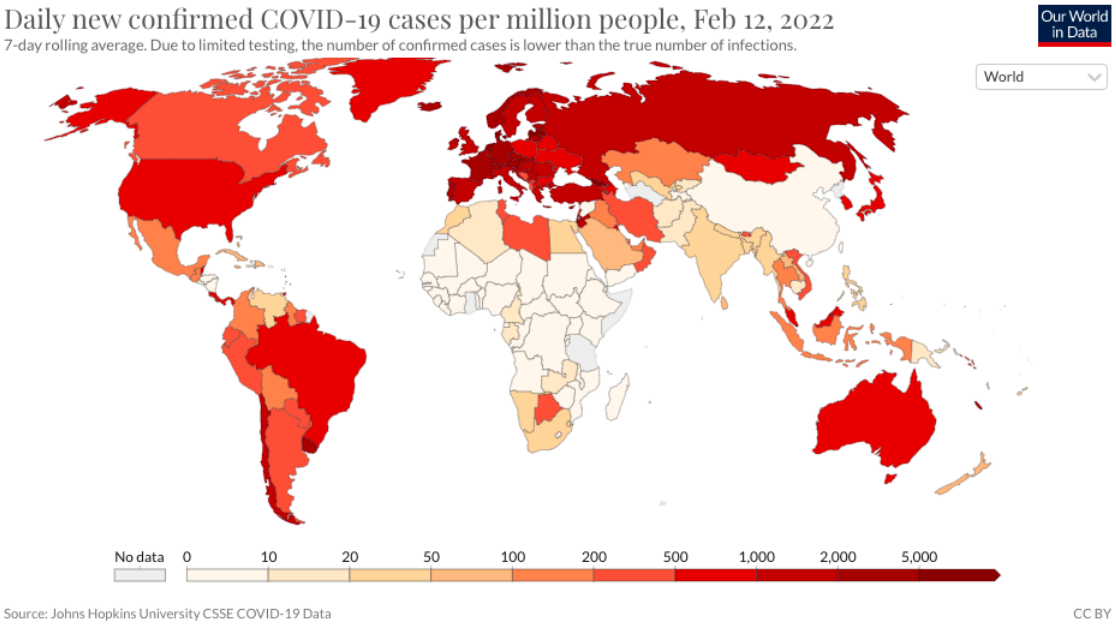
Εικόνα 12: Απεικόνιση νέων κρουσμάτων και θανάτων Covid-19 ανά εκατομμύριο για κάθε χώρα σε μορφή πίνακα [22].



Εικόνα 13: Σύγκριση νέων κρουσμάτων και θανάτων Covid-19 μεταξύ Ελλάδας και Ισπανίας κατά τη διάρκεια της πανδημίας [22].



Εικόνα 14: Σύγκριση ημερήσιου αριθμού κρουσμάτων για πέντε διαφορετικές χώρες [22].



Εικόνα 15: Ημερήσιος αριθμός κρουσμάτων σε παγκόσμια κλίμακα για επιλεγμένη ημερομηνία [22].

3.2.2 Μετακίνηση

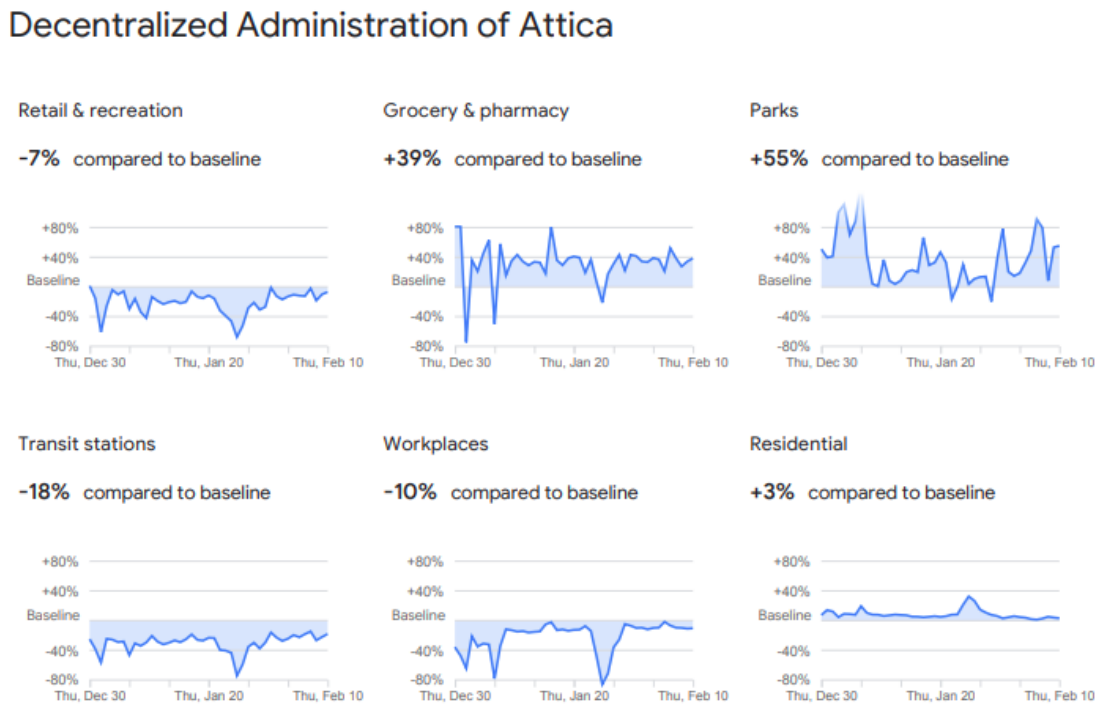
Το επόμενο σημαντικό σύνολο δεδομένων που χρησιμοποιήθηκε για τους σκοπούς της παρούσας διπλωματικής εργασίας δημιουργήθηκε από την Google [23] και σχετίζεται με τις μετακινήσεις των ανθρώπων κατά την περίοδο της πανδημίας. Η δομή τους παρουσιάζεται στον Πίνακας 2 Πιο συγκεκριμένα, οι μετακινήσεις χωρίζονται σε έξι βασικές κατηγορίες, σύμφωνα με τις προδιαγραφές του του συγκεκριμένου συνόλου: φαρμακεία και παντοπωλεία, πάρκα, χώρους εργασίας, κατοικίες, σταθμούς διέλευσης και χώρους αναψυχής.

Πίνακας 2: Δεδομένα που σχετίζονται με τη μετακίνηση του πληθυσμού.

Mobility	grocery_and_pharmacy_percent_change_from_baseline
	parks_percent_change_from_baseline
	workplaces_percent_change_from_baseline
	residential_percent_change_from_baseline

Όλα τα δεδομένα αυτού του συνόλου μετριούνται συγκριτικά με τη βάση (baseline) της κάθε κατηγορίας. Η βάση αυτή προέρχεται από τον υπολογισμό της μέσης τιμής για την αντίστοιχη μέρα της εβδομάδας κατά την περίοδο από 03 Ιανουαρίου έως 06 Φεβρουαρίου 2020.

Στην Εικόνα 16 παρουσιάζονται τα δεδομένα μετακίνησης που αφορούν την περιοχή της Αττικής για ένα συγκεκριμένο χρονικό διάστημα της επιλογής του χρήστη που μπορεί να ξεκινάει στις αρχές του έτους 2020 και φτάνει έως και το παρόν. Στα πλαίσια της παρούσας εργασίας εξετάστηκε μόνο η περίοδος που αφορά το έτος 2020.



Εικόνα 16: Δεδομένα μετακίνησης ανθρώπων στην περιοχή της Αττικής για έξι διαφορετικές κατηγορίες [23].

3.2.3 Ατμοσφαιρικά Δεδομένα

Το σύνολο των ατμοσφαιρικών δεδομένων προέρχεται από το project «World air quality index» [24], του οποίου η δομή φαίνεται στον Πίνακας 3, ξεκίνησε το 2007 και αποστολή του είναι να ενισχύει την επίγνωση των πολιτών σχετικά με τη μόλυνση του αέρα και να παρέχει ομογενή και παγκόσμια πληροφορία για την ποιότητα του αέρα. Τα δεδομένα παράγονται σε ημερήσια βάση και προέρχονται από εγκατεστημένους σταθμούς στις διάφορες πόλεις του κόσμου. Χρησιμοποιώντας τα δεδομένα που συλλέγονται καθημερινά, δημιουργείται ένας ημερήσιος δείκτης για κάθε περιοχή, η τιμή του οποίου καθορίζει την ποιότητα του αέρα.

Πίνακας 3: Δεδομένα που σχετίζονται με τη μόλυνση του αέρα.

Air Pollution	PM2.5 concentration
	PM10 concetration

	NO2 concetration
	O3 concetration
	Pressure
	Wind
	Temperature

Η Εικόνα 17 παρουσιάζει τον ημερήσιο δείκτη ποιότητας αέρα κάθε συμβεβλημένης περιοχής που εμφανίζεται στο χάρτη. Η κλίμακα ποιότητας του αέρα διαχωρίζεται στις εξής κατηγορίες, οι οποίες αναλύονται εκτενέστερα στον Πίνακας 4:

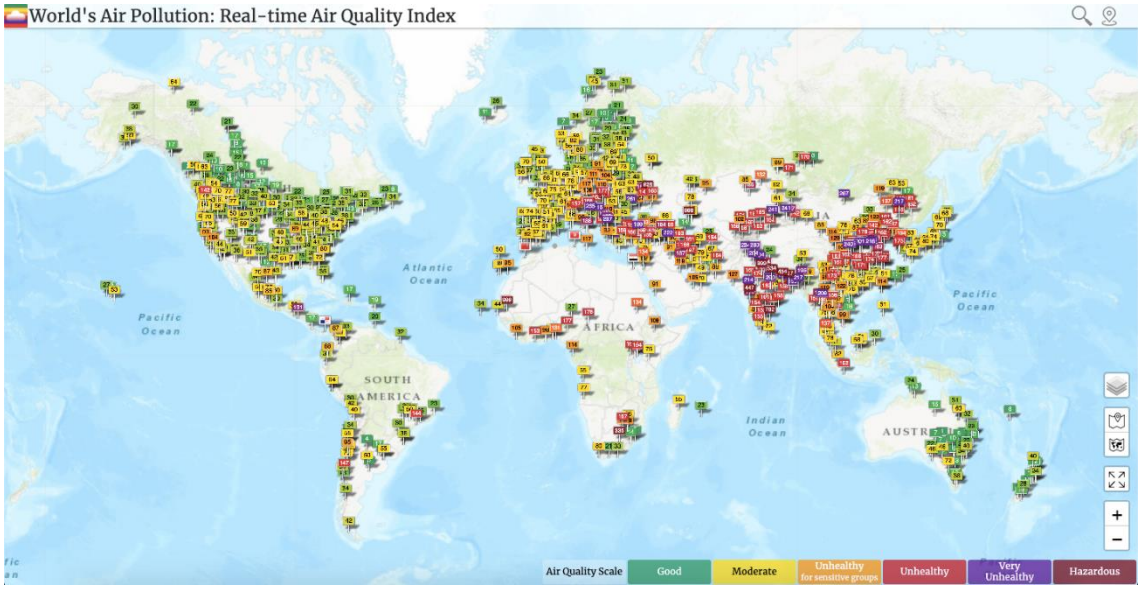
- Καλό
- Μέτριο
- Μη υγιές για ευάλωτες κατηγορίες
- Μη υγιές
- Εξαιρετικά μη υγιές
- Επικίνδυνο

Μία περιοχή μπορεί να ενταχθεί στις επόμενες κατηγορίες (Πίνακας 4) ανάλογα με την τιμή του ημερήσιου δείκτη της.

Πίνακας 4: Επιπλοκές υγείας ανάλογα με το επίπεδο μόλυνσης του αέρα.

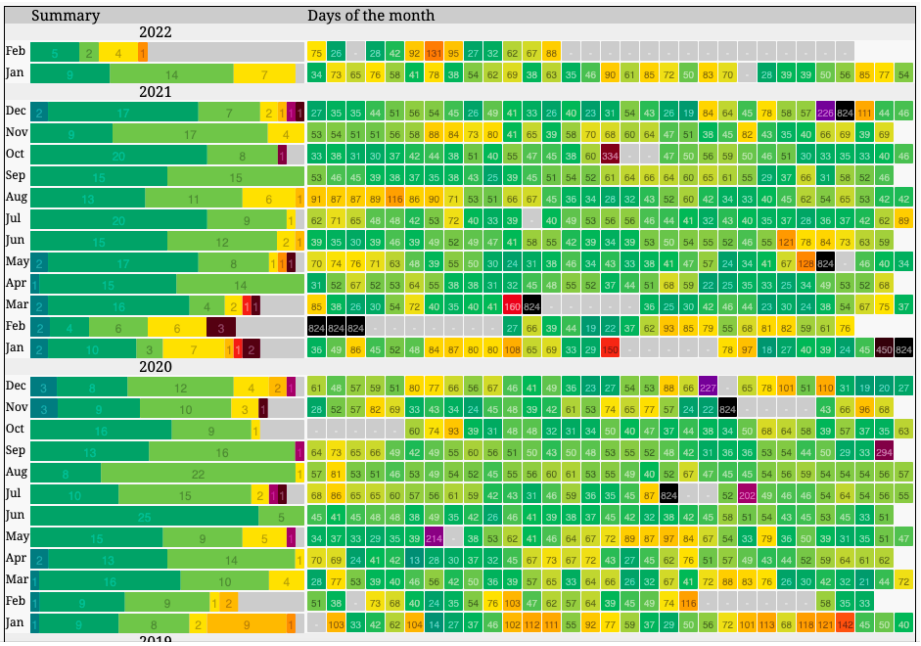
Δείκτης ποιότητας αέρα	Επίπεδο μόλυνσης αέρα	Επιπλοκές υγείας
0 – 50	Καλό	Η ποιότητα του αέρα θεωρείται ικανοποιητική και η μόλυνση εμφανίζει λίγο ή καθόλου κίνδυνο.
51 - 100	Μέτριο	Η ποιότητα αέρα είναι αποδεκτή. Όμως, για κάποιους ρύπους υπάρχει μία μικρή ανησυχία, όσον αφορά συγκεκριμένη μερίδα πληθυσμού που είναι ευαίσθητη στη μολυνση του αέρα.

101 – 150	Μη υγιές για ευάλωτες κατηγορίες	Τα μέλη των ευαίσθητων κατηγοριών μπορεί να αντιμετωπίσουν προβλήματα με την υγεία τους. Παρ' όλα αυτά, ο γενικός πληθυσμός δεν είναι πιθανό να επηρεαστεί σημαντικά.
151 – 200	Μη υγιές	Όλοι μπορεί να αρχίσουν να αντιμετωπίζουν προβλήματα υγείας και κατ' επέκταση τα μέλη των ευαίσθητων ομάδων μπορεί να αντιμετωπίσουν πολύ σοβαρότερα προβλήματα υγείας.
201 – 300	Εξαιρετικά μη υγιές	Ολόκληρος ο πληθυσμός είναι πιθανό να επηρεαστεί με προβλήματα υγείας.
300+	Επικίνδυνο	Συναγερμός υγείας. Όλοι μπορεί να αντιμετωπίσουν πιο σοβαρά προβλήματα υγείας.



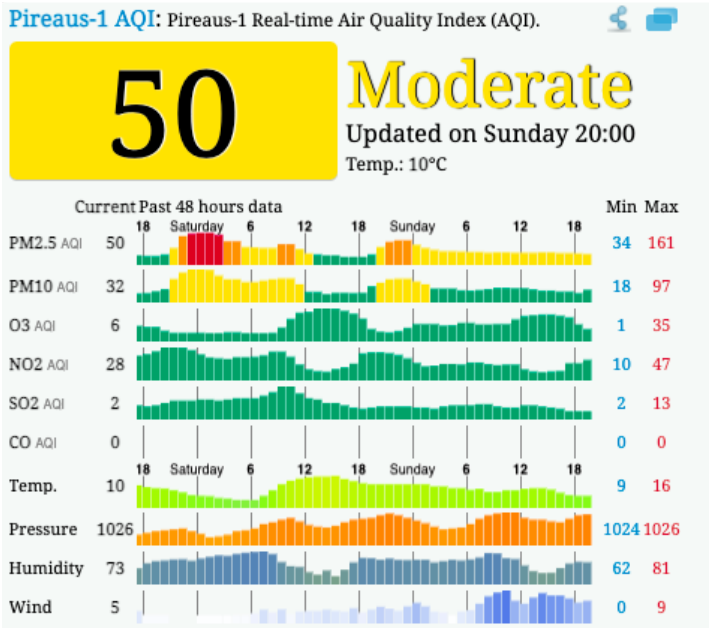
Εικόνα 17: Δείκτης ποιότητας αέρα σε παγκόσμια κλίμακα [24].

Στην Εικόνα 18 παρουσιάζονται οι ημερήσιες παρατηρήσεις επιπέδου ποιότητας αέρα για το χρονικό διάστημα από 01/01/2020 έως την τρέχουσα ημερομηνία του 2022.



Εικόνα 18: Παρατήρηση επιπέδου ποιότητας αέρα μίας συγκεκριμένης περιοχής(Πειραιάς) για τα έτη 2020-2022 [24].

Στην Εικόνα 19 παρουσιάζονται οι ωριαίες παρατηρήσεις επιπέδου ποιότητας αέρα για το τελευταίο 48ωρο της τρέχουσας ημερομηνίας για την περιοχή του Πειραιά.



Εικόνα 19: Ωριαία παρατήρηση δεικτών ποιότητας αέρα για την περιοχή του Πειραιά [24].

3.2.4 Νοσοκομεία

Τα δεδομένα που σχετίζονται με τον αριθμό νοσοκομείων που υπάρχουν σε μία περιοχή υπολογίζονται χρησιμοποιώντας την γεωγραφική θέση της περιοχής αυτής και της επιθυμητής ακτίνας αναζήτησης. Τα δεδομένα προέρχονται από πηγές των Wikidata και OpenStreetMap [27], [29].

3.2.5 Χώροι αναψυχής/ψυχαγωγίας

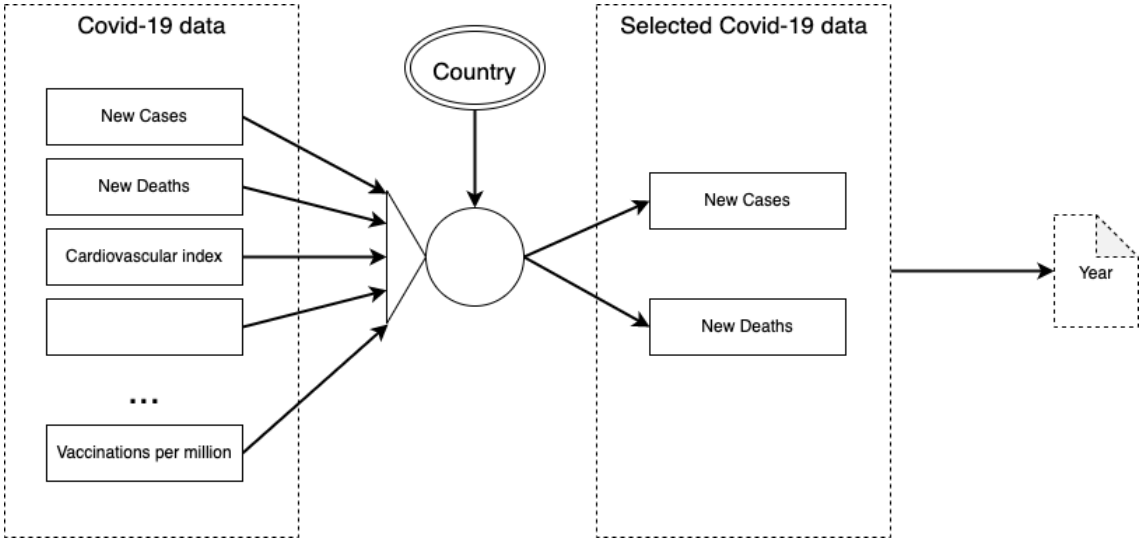
Οι χώροι αναψυχής σχετίζονται κυρίως με αθλητικά κέντρα και πιο συγκεκριμένα γήπεδα. Για την ακρίβεια, με χρήση των κατάλληλων ερωτημάτων – queries εξάγεται η αντίστοιχη πληροφορία που σχετίζεται με τον αριθμό των γηπέδων που βρίσκονται σε μία περιοχή, που ορίζεται από την γεωγραφική της θέση και την επιθυμητή ακτίνα αναζήτησης. Τα δεδομένα προέρχονται από πηγές των Wikidata και OpenStreetMap [27], [29].

3.3 Προετοιμασία δεδομένων

Από τη συλλογή των δεδομένων σχετικά με την παρακολούθηση της πανδημίας Covid-19 προέκυψε ένα αρχείο σε μορφή .csv, το οποίο περιλαμβάνει όλα τα σχετικά στατιστικά δεδομένα για κάθε χώρα του κόσμου. Τα δεδομένα αυτά είναι τα εξής:

- Επιβεβαιωμένοι θάνατοι
- Επιβεβαιωμένα κρούσματα
- Τεστ ανά κρούσμα
- Δείκτης αυστηρότητας
- Κρούσματα
- Θάνατοι
- Εκτεταμένη θνησιμότητα

Για τους σκοπούς της παρούσας διπλωματικής επιλέχθηκε μόνο η πληροφορία που σχετίζεται με τα ημερήσια κρούσματα και θανάτους, χρησιμοποιώντας την επιλογή της εκάστοτε χώρας, όπως παρουσιάζεται στην Εικόνα 20.

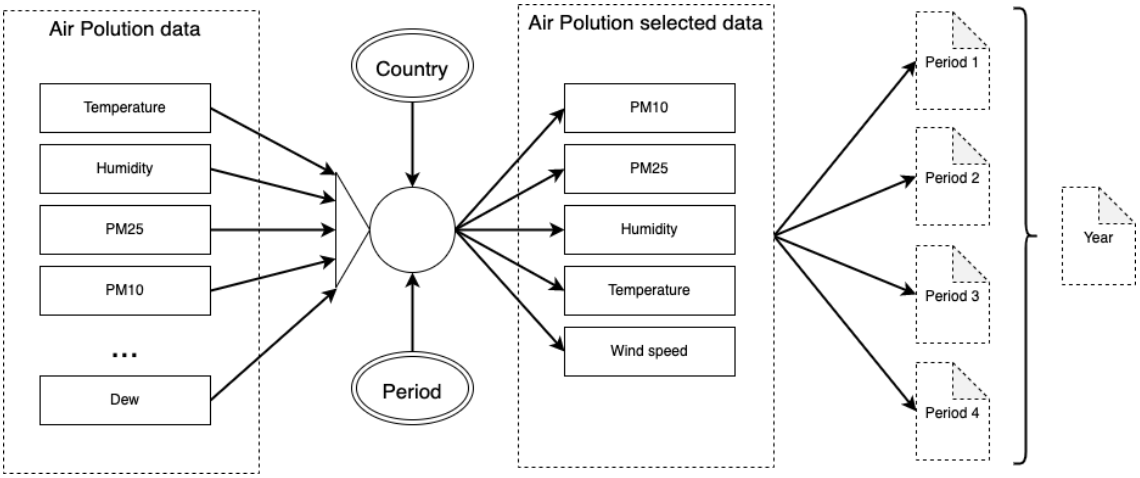


Εικόνα 20: Επιλεγμένα δεδομένα που χρησιμοποιήθηκαν σχετικά με Covid-19.

Από τη συλλογή των δεδομένων σχετικά με την ατμοσφαιρική μόλυνση προέκυψαν πολλαπλά αρχεία σε μορφή .csv, λόγω περιορισμών της πλατφόρμας σχετικά με την περίοδο συλλογής των δεδομένων. Πιο συγκεκριμένα, μετά τον ορισμό της χώρας ενδιαφέροντος, προκύπτουν τέσσερα csv αρχεία που αντιστοιχούν στα τέσσερα ακολουθιακά 3μηνα του έτους 2020. Αυτά τα τέσσερα αρχεία ενσωματώθηκαν σε ένα αρχείο, που περιλαμβάνει τα εν λόγω δεδομένα και παρουσιάζεται στην Εικόνα 21.

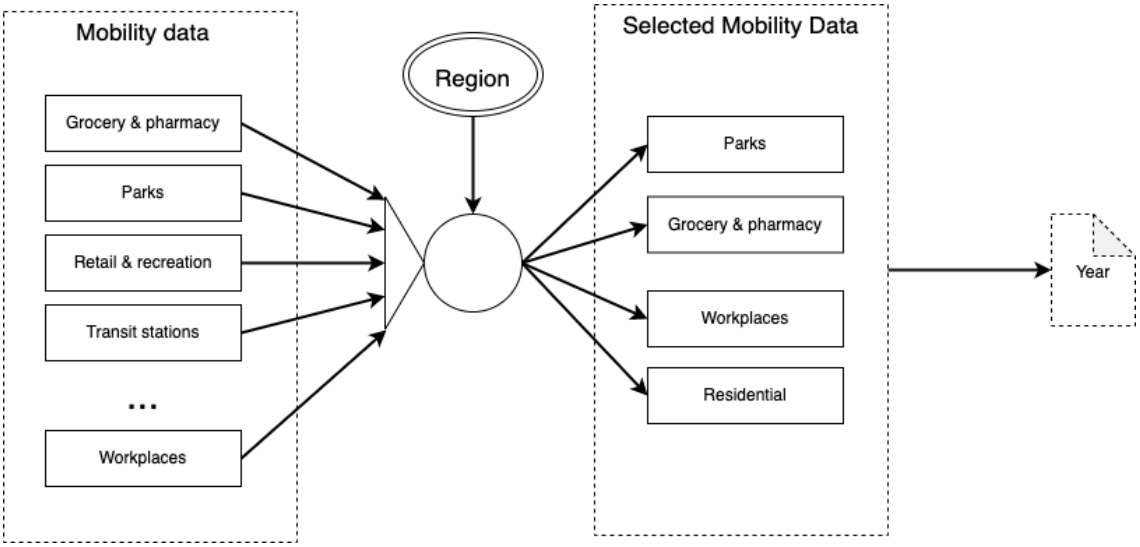
Από αυτά επιλέχθηκαν οι παρακάτω δείκτες για περαιτέρω ανάλυση:

- Συγκέντρωση σωματιδίων PM25
- Συγκέντρωση σωματιδίων PM10
- Όζον O3
- Υγρασία



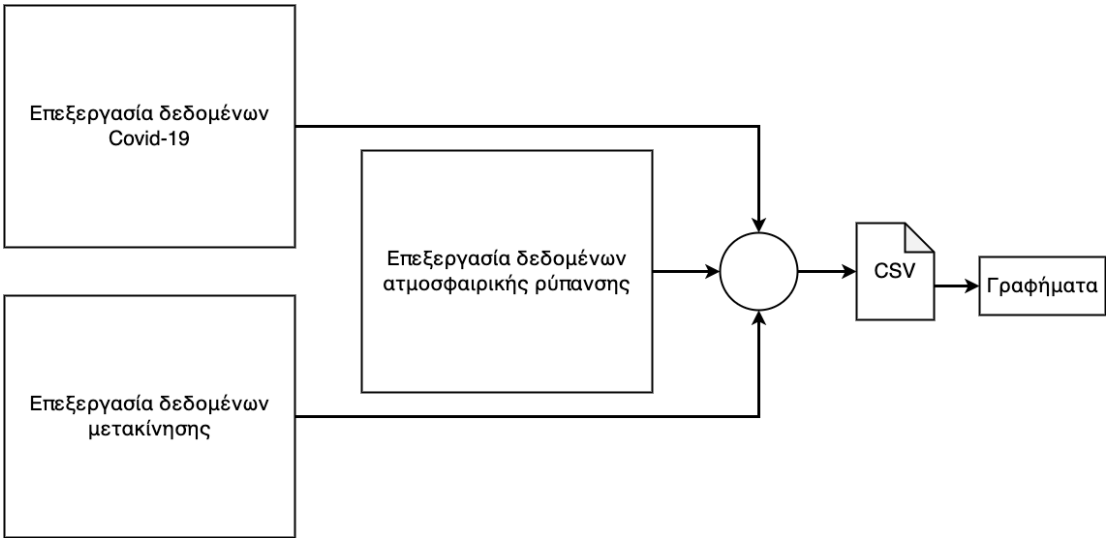
Εικόνα 21: Επιλεγμένα δεδομένα ατμοσφαιρικής μόλυνσης.

Από τη συλλογή των δεδομένων σχετικά με την μετακίνηση προέκυψε ένα αρχείο σε μορφή .csv για κάθε πόλη ενδιαφέροντος για το έτος 2020, όπως απεικονίζεται στην Εικόνα 22.



Εικόνα 22: Επιλεγμένα δεδομένα μετακίνησης.

Η συνολική διαδικασία κατασκευής δεδομένων απεικονίζεται στην Εικόνα 23, όπου ένα τελικό αρχείο προέκυψε από το συνδυασμό όλων των παραπάνω .csv αρχείων.



Εικόνα 23: Συνολική διαδικασία κατασκευής επιλεγμένων δεδομένων.

3.4 Σχεδίαση και ανάπτυξη οντολογίας

Για τη δημιουργία της οντολογίας χρησιμοποιήθηκε το εργαλείο Protégé [14], το οποίο αποτελεί ένα εργαλείο ανοιχτού κώδικα που προσφέρει ένα περιβάλλον ανάπτυξης και επεξεργασίας οντολογιών σε OWL.

Βασικά στοιχεία μίας οντολογίας αποτελούν:

● Κλάσεις (classes)

● Συσχετίσεις (relations)

● Λογικούς τελεστές (Operators)

● Αξιώματα (axioms)

● Στιγμιότυπα (instances - individuals)

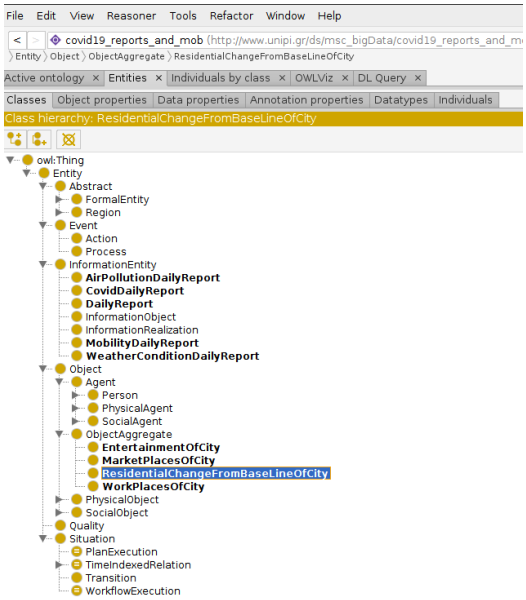
Οι κλάσεις είναι σύνολα στιγμιοτύπων με κοινές ιδιότητες/χαρακτηριστικά και μπορούν να χρησιμοποιηθούν για κατηγοριοποίηση, τυποποίηση, ταξινόμηση, κλπ. Ένα μέλος μίας κλάσης εξαρτάται από τη λογική περιγραφή της και όχι από το όνομα. Οι κλάσεις δεν είναι απαραίτητο να έχουν όνομα, μπορεί να είναι λογικές εκφράσεις, όπως «τα πράγματα που έχουν χρώμα μπλε». Επίσης, μία κλάση πρέπει να περιγράφεται με τέτοιο τρόπο, ώστε να μπορεί να περιλαμβάνει τα στιγμιότυπα (εκτός αν αναφερόμαστε σε μία άδεια κλάση). Για παράδειγμα, σε μία οντολογία που αφορά ένα νοσοκομείο, ο γιατρός και ο ασθενής αποτελούν κλάσεις.

Το Protégé δίνει τη δυνατότητα στο χρήστη να ορίσει τις κλάσεις και την μεταξύ τους ιεραρχία, όπως απεικονίζεται στην Εικόνα 24. Οι κλάσεις που δημιουργήθηκαν για την οργάνωση των δεδομένων αυτής της εργασίας φαίνονται με έντονη γραμματοσειρά στο παράθυρο ιεραρχίας των κλάσεων. Αναλυτικότερα:

- AirpollutionDailyReport - περιλαμβάνει στιγμιότυπα που αφορούν ημερήσιες αναφορές μόλυνσης αέρα,
- CovidDailyReport - σχετίζεται με τα νέα κρούσματα και θανάτους από Covid-19 σε ημερίσια βάση,
- DailyReport - είναι μία γενική κλάση για μία πιο δομημένη περιγραφή της ημερομηνίας ανάκτησης της κάθε πληροφορίας,
- MobilityDailyReport - αναφέρεται στα δεδομένα που σχετίζεται με την μετακίνηση των ανθρώπων και αναλύεται εκτενέστερα στις υποκλάσεις EntertainmentOfCity,

MarketPlacesOfCity, ResidentialChangeFromBaselineOfCity και WorkPlacesOfCity,

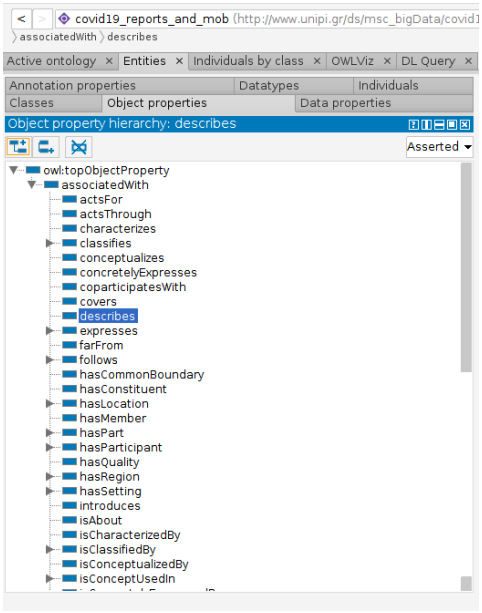
- WeatherConditionDailyReport – περιλαμβάνει στιγμιότυπα ημερήσιων αναφορών καιρικών συνθηκών.



Εικόνα 24: Η δομή των κλάσεων της οντολογίας.

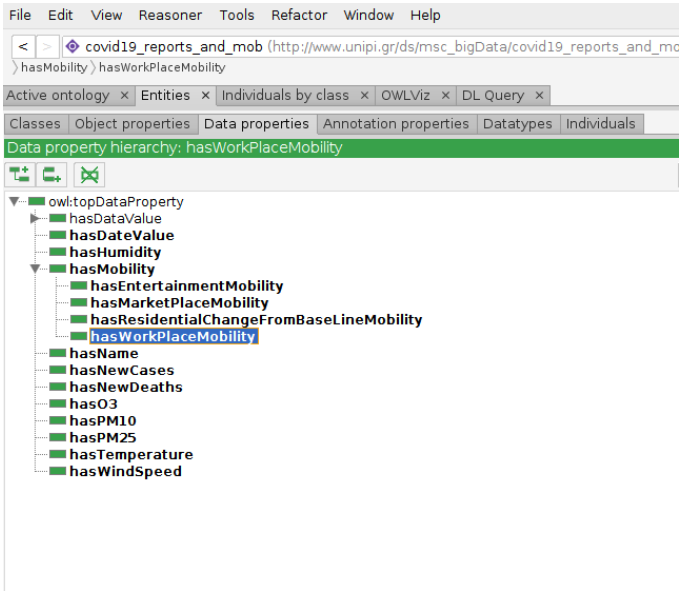
Αφού δηλωθούν οι κλάσεις της οντολογίας, δηλώνονται και οι συσχετίσεις των στιγμιοτύπων (ObjectProperties). Όπως παρουσιάζεται και στην Εικόνα 25, για τον ορισμό τους είναι απαραίτητο να οριστούν τουλάχιστον ο τύπος των υποκειμένων και αντικειμένων κάθε συσχέτισης. Τα dataTypeProperties συσχετίζουν τα στιγμιότυπα στις τιμές δεδομένων.

Στην Εικόνα 25 φαίνονται πρωτότυπα σχέσεων που μπορούν να χρησιμοποιηθούν κατά την ανάπτυξη μίας οντολογίας και φυσικά κατά την διαδικασία κατασκευής των ζητούμενων ερωτημάτων - queries σε SPARQL για την ανάκτηση της ζητούμενης πληροφορίας. Στα πλαίσια της παρούσας εργασίας επιλέχθηκαν μερικές από τις σχέσεις αυτές όπως η hasLocation και η hasRegion.



Εικόνα 25: Η δομή των σχέσεων των αντικειμένων και κλάσεων της οντολογίας

Έπειτα, πρέπει να προσδιοριστούν τα χαρακτηριστικά των αντικειμένων και κλάσεων, δηλαδή τα dataProperties, όπως φαίνεται στην Εικόνα 26.

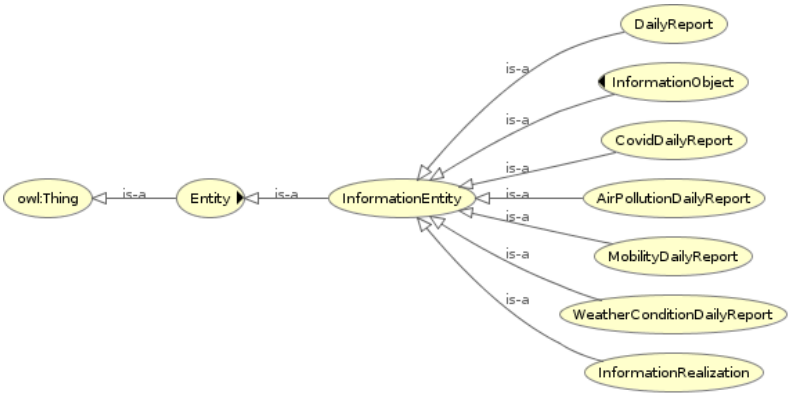


Εικόνα 26: Τα χαρακτηριστικά των αντικειμένων και κλάσεων της οντολογίας.

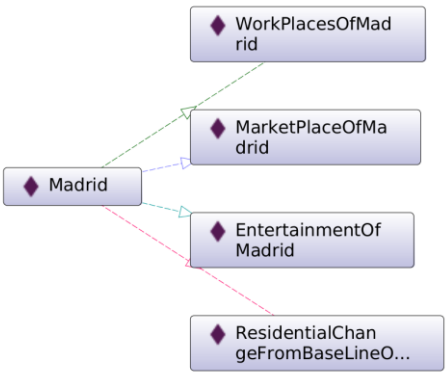
Στη συνέχεια, γίνεται η δημιουργία των στιγμιοτύπων (Individuals) της οντολογίας, που ουσιαστικά αναφέρονται σε συγκεκριμένα υποκείμενα/αντικείμενα που περιέχουν τις διασυνδέσεις και λειτουργίες της εκάστοτε κλάσης. Στις Εικόνες 27 – 32 παρουσιάζονται τα δομικά στοιχεία της οντολογίας που έχει δημιουργηθεί μέσω του εργαλείου απεικόνισης του Protégé. Η οργάνωση των δεδομένων, που συλλέχτηκαν για τους σκοπούς της παρούσας εργασίας, εντάχθηκε στην κλάση InformationEntity. Με άλλα λόγια, οι κλάσεις που δημιουργήθηκαν και αναφέρονται στα δεδομένα σχετικά με Covid-19, το περιβάλλον και την μετακίνηση των ανθρώπων, θεωρούνται υποκλάσεις της οντότητας InformationEntity. Ταυτόχρονα, οι υπόλοιπες οντότητες όπως η Event, η Situation κτλ αντιπροσωπεύουν άλλα είδη δεδομένων που μπορεί να συλλεχθούν στο μέλλον και να ενταχθούν στην γενικευμένη οντολογία.

Εικόνα 27: Η γενική δομή της οντολογίας.

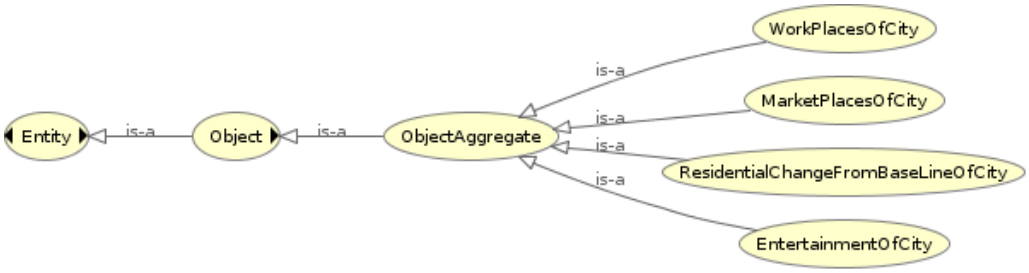
Στην Εικόνα 28 παρουσιάζεται ξεχωριστά η κλάση InformationEntity που περιλαμβάνει όλες τις σχετικές με τα δεδομένα κλάσεις, οι οποίες συνδέονται με σχέσεις τύπου ISA, καθώς αποτελούν υποκλάσεις. Οι κλάσεις που σχετίζονται με τα δεδομένα μετακίνησης των ανθρώπων εξειδίκευονται σε υποκλάσεις ανάλογα με τον τύπο της κάθε μετακίνησης, όπως φαίνεται στην Εικόνα 30. Κατ' επέκταση κάθε τύπος μετακίνησης αναπαράγεται για καθεμία από τις πόλεις-πιλότους που αποτελούν individuals στο τελικό σχήμα. Ένα στιγμιότυπο που αναπαριστά μια πόλη (Madrid, Εικόνα 29) συσχετίζεται με στιγμιότυπα που αναπαριστούν χώρους δραστηριότητας (workplacesOfMadrid, marketPlacesOfMadrid, κλπ). Αντίστοιχα δημιουργούνται οι συσχετίσεις και για τις υπόλοιπες πόλεις-πιλότους, π.χ. το στιγμιότυπο Athens συσχετίζεται με το workplacesOfAthens, κλπ. Μία τέτοια σχέση μπορεί να χαρακτηριστεί ως σχέση Many-to-Many καθώς κάθε πόλη σχετίζεται με κάθε τύπο μετακίνησης και κάθε τύπος μετακίνησης συνδέεται με κάθε πόλη. Βάσει αυτού του χαρακτηριστικού της οντολογίας επιλέχθηκε η συγκεκριμένη δομή των Εικόνων Εικόνα 29, Εικόνα 31.



Εικόνα 28: Οι κλάσεις που σχετίζονται με τα συλλεχθέντα δεδομένα.

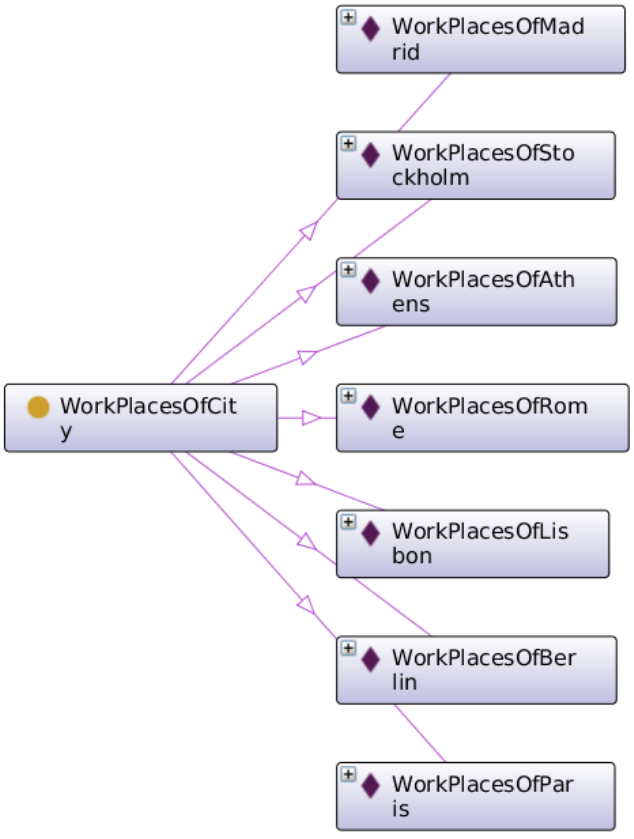


Εικόνα 29: Μοντελοποίηση των δεδομένων μετακίνησης πολιτών στην πόλη της Μαδρίτης



Εικόνα 30: Οι κλάσεις που σχετίζονται με τα δεδομένα μετακίνησης στην οντολογία.

Μία διαφορετική οπτική δείχνει η Εικόνα 31 όπου παρουσιάζεται ο τρόπος με τον οποίο προκύπτουν τα στιγμιότυπα της κλάσης workplacesOfCity για κάθε πόλη – πιλότο που σχετίζονται με δεδομένα μετακίνησης των ανθρώπων προς τους χώρους εργασίας τους.



Εικόνα 31: Μοντελοποίηση του δείκτη παραμονής στην κατοικία και πώς συσχετίζεται με κάθε πόλη πιλότο.

3.5 Μετασχηματισμός δεδομένων σε RDF

Μετά τη δημιουργία της οντολογίας ακολούθησε ο μετασχηματισμός των δεδομένων από CSV αρχεία σε RDF, ο οποίος πραγματοποιήθηκε χρησιμοποιώντας το εργαλείο RDF-GEN [28] με προσθήκη των κατάλληλων κανόνων και τη δημιουργία του αντίστοιχου αρχείου ρυθμίσεων. Στην Εικόνα 33 παρουσιάζεται η ροή των δεδομένων μέσω του μηχανισμού παραγωγής τριπλετών (Triple Generator), ο οποίος δέχεται για είσοδο τα αρχεία CSV και δημιουργεί τις τριπλέτες εξόδου. Τα αρχεία εξόδου έχουν κατάληξη .ttl και περιλαμβάνουν τα δεδομένα σε μορφή τριπλετών που ακολουθούν τους κανόνες που έχουν οριστεί στα αρχεία ρυθμίσεων.

Το διάνυσμα που περιέχει τις μεταβλητές των δεδομένων εισόδου (Vector of Variables, Εικόνα 33) δημιουργείται στα αρχεία με κατάληξη .rdf/.xml, που χρησιμοποιούνται ως αρχεία διαμόρφωσης (configuration files) από το εργαλείο RDF-Gen. Πιο αναλυτικά τα αρχεία αυτά εμπεριέχουν όλη την γενική πληροφορία σχετικά με την τοποθεσία όλων των απαραίτητων αρχείων, τον ορισμό γενικών παραμέτρων και φυσικά τα ονόματα των μεταβλητών από τα δεδομένα εισόδου. Πιο συγκεκριμένα, ένα αρχείο ρυθμίσεων (σε μορφή rdf/xml) περιέχει τα παρακάτω χαρακτηριστικά:

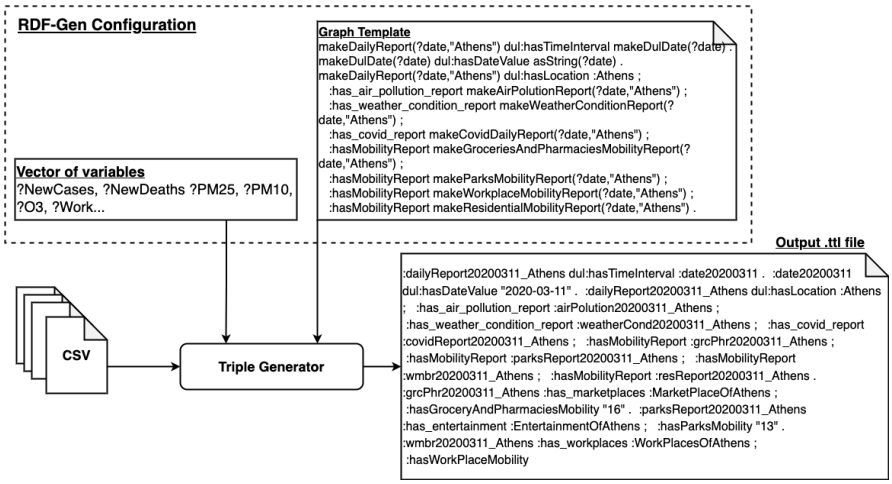
- Datasource - όπου ορίζεται η ακριβής τοποθεσία του αρχείου εισόδου στο σύστημα αρχείων του υπολογιστή,
- Connector - όπου ορίζεται ο τύπος αρχείου εισόδου, CSV στην περίπτωση της παρούσας διπλωματικής εργασίας,
- Template - αφορά στην δήλωση της τοποθεσίας του αρχείου που σχετίζεται με το μοτίβο μετασχηματισμού των δεδομένων σε τριπλέτες,
- TemplateVariables - περιλαμβάνει όλες τις μεταβλητές του template που θα χρησιμοποιηθούν για την παραγωγή των τριπλετών,
- InputVariables - περιλαμβάνει τις μεταβλητές (π.χ. στήλες στα αρχεία CSV) που εντοπίζονται στο αρχείο εισόδου,
- Delimiter - όπου δηλώνεται ο χαρακτήρας που χρησιμοποιήθηκε για τον διαχωρισμό των τιμών που περιλαμβάνονται στα CSV αρχεία,
- Prefix - πεδίο που δηλώνει την τοποθεσία του αρχείου όπου δίνονται τα προθέματα των URIs του αρχείου εξόδου..

Ένα ολοκληρωμένο αρχείο ρυθμίσεων παρουσιάζεται στην Εικόνα 32.

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dcf="http://www.datacron-project.eu/RDFGen_conf#"
>

  <!-- the absolute path to data source, this can be a folder to be recursively processed -->
  <dcf:DataSource rdf:about="file:patsimas/data/Athens.csv">

    <!-- Required: specifies the connector to be used -->
    <dcf:connector>csv</dcf:connector>

    <!-- Required: the template to be used -->
    <dcf:template>patsimas/patsimas_Athens.q</dcf:template>

    <!-- Required: the variables to be used for the source in the order specified -->
    <dcf:TemplateVariables>?id,?date,?wind_speed,?temperature,?pm25,?pm10,?o3,?humidity,?new_cases,
      ?new_deaths,?parks_percent_change_from_baseline,
      ?grocery_and_pharmacy_percent_change_from_baseline,
      ?workplaces_percent_change_from_baseline,?residential_percent_change_from_baseline</dcf:TemplateVariables>

    <dcf:inputVariables>0,1,2,3,4,5,6,7,8,9,10,11,12,13,14</dcf:inputVariables>

    <!-- Required (for csv and 7z data sources): the delimiter used in the source-->
    <dcf:delimiter>,</dcf:delimiter>
    <dcf:skip>1</dcf:skip>

    <!--optional: if provided the generated triples will be provided at this path -->
    <dcf:output>Athens_output.ttl</dcf:output>

    <!--optional: if provided the prefix will be added on top of the output.ttl ONCE AND WHEN RDF-Gen IS FINISHED -->
    <dcf:prefix>file://home/giorgos/workspace/RDF_Gen_Flink/prefix.ttl</dcf:prefix>


  </dcf:DataSource>

</rdf:RDF>
```

Εικόνα 32: Το αρχείο ρυθμίσεων για τον μετασχηματισμό των CSV δεδομένων που αφορούν την Αθήνα.

Ο Πίνακας 5 παρουσιάζει τα δεδομένα σε μορφή CSV που προέκυψαν μετά την επεξεργασία των αρχικών δεδομένων από τις πηγές. Η Εικόνα 34 δείχνει ένα από τα αρχεία .q – πρότυπα γράφου (graph template) που αντιπροσωπεύουν την γραμματική της σχετιζόμενης οντολογίας.



Εικόνα 33: Περιγραφή του τρόπου λειτουργίας του εργαλείου RDF-Gen.

Πίνακας 5: Παράδειγμα δεδομένων που περιλαμβάνονται σε csv αρχείο.

date	wind_speed	wind_gust	temperature	co	so2	pressure	pm25	pm10	o3	no2	humidity	dew
2019-12-30	2	-200	5	0.1	2.1	1013.3	21	7	10.2	6	86	-200
2019-12-31	2	-200	5	0.1	3.1	1013.3	34	8	16.7	6.9	86	-200
2020-01-01	-200	-200	-200	0.1	3.1	-200	57	19	23.2	7.8	-200	-200
2020-01-02	7	12.8	6.6	0.1	2.6	1024	34	11	27.7	6.9	49	-2
2020-01-03	3.6	7.9	7	0.1	4.1	1020.9	46	13	25.6	9.2	50.5	-1
2020-01-04	1	3.1	9	0.1	4.6	1016.1	50	14	27.7	14.7	44.8	-2.5
2020-01-05	3.6	8.5	8.3	0.1	3.1	1006.4	50	12	24	9.6	69.6	3.5
2020-01-06	7.2	15.4	5	0.1	3.1	1007.3	17	5	22.4	6.4	85.8	2.5
2020-01-07	5.1	12.6	5.5	0.1	2.6	1016.7	30	6	25.6	4.6	74.7	-1
2020-01-08	3.8	10.1	6.5	0.1	4.1	1021.1	30	11	27.7	6	55.5	-3
2020-01-09	2.8	7.9	7.2	0.1	6.6	1022.8	42	14	28.1	10.6	53.5	-2
2020-01-10	1.5	3.6	7.5	0.1	6.6	1021	59	20	15.1	22.9	54.8	-3
2020-01-11	1.2	3.6	8.3	0.1	2.6	1021	76	30	15.9	20.2	64.9	4
2020-01-12	1.5	3.8	9.7	0.1	5.6	1022.5	76	26	22.4	10.6	71	4

Με άλλα λόγια, σε αυτά τα αρχεία περιλαμβάνονται όλοι οι απαραίτητοι κανόνες που χρησιμοποιεί το εργαλείο RDF_Gen για να δημιουργήσει τις ζητούμενες τριπλέτες και να δώσει για έξοδο τα αντίστοιχα .ttl αρχεία όπως φαίνεται στην Εικόνα 33.

```
makeDailyReport(?date,"Athens") dul:hasTimeInterval makeDulDate(?date) .
makeDulDate(?date) dul:hasDateValue asString(?date) .
makeDailyReport(?date,"Athens") dul:hasLocation :Athens ;
   :has_air_pollution_report makeAirPolutionReport(?date,"Athens") ;
   :has_weather_condition_report makeWeatherConditionReport(?date,"Athens") ;
   :has_covid_report makeCovidDailyReport(?date,"Athens") ;
   :hasMobilityReport makeGroceriesAndPharmaciesMobilityReport(?date,"Athens") ;
   :hasMobilityReport makeParksMobilityReport(?date,"Athens") ;
   :hasMobilityReport makeWorkplaceMobilityReport(?date,"Athens") ;
   :hasMobilityReport makeResidentialMobilityReport(?date,"Athens") .


makeGroceriesAndPharmaciesMobilityReport(?date,":Athens") :has_marketplaces :MarketPlaceOfAthens ;
   :hasGroceryAndPharmaciesMobility asString(?grocery_and_pharmacy_percent_change_from_baseline) .
makeParksMobilityReport(?date,":Athens") :has_entertainment :EntertainmentOfAthens ;
   :hasParksMobility asString(?parks_percent_change_from_baseline) .
makeWorkplaceMobilityReport(?date,"Athens") :has_workplaces :WorkPlacesOfAthens ;
   :hasWorkPlaceMobility asString(?workplaces_percent_change_from_baseline) .
makeResidentialMobilityReport(?date,":Athens") :has_residential_change_from_baseline
   :ResidentialChangeFromBaseLineOfAthens ;
   :hasResidentialMobility asString(?residential_percent_change_from_baseline) .

makeAirPolutionReport(?date,"Athens")  dul:hasLocation :Athens ;
   :hasPM25 asFloat(?pm25) ;
   :hasO3 asFloat(?o3) ;
   :hasPM10 asFloat(?pm10) .

makeWeatherConditionReport(?date,"Athens")  dul:hasLocation :Athens ;
   :hasWindSpeed asFloat(?wind_speed) ;
   :hasHumidity asFloat(?humidity) ;
   :hasTemperature asFloat(?temperature) .

makeCovidDailyReport(?date,"Athens")  dul:hasLocation :Athens ;
   :hasNewCases asInteger(?new_cases) ;
   :hasNewDeaths asInteger(?new_deaths) .

# we also support constants in functions: e.g. test(?date,"hello").
```

Εικόνα 34: Παράδειγμα .q αρχείου που περιγράφει το πρότυπο του γράφου της οντολογίας.

4 Αποτελέσματα

Μετά τη συλλογή, κατασκευή και κατάλληλη ομαδοποίηση των δεδομένων, χρησιμοποιήθηκαν οι γνώσεις σχετικά με την οργάνωση των δεδομένων κατά τις αρχές του σημασιολογικού ιστού, ώστε να σχεδιαστεί και να υλοποιηθεί η επιθυμητή οντολογία, με την χρήση των δυνατοτήτων του εργαλείου Protege. Στην συνέχεια, με την αξιοποίηση των χαρακτηριστικών της οντολογίας και συνεπώς των αντίστοιχων κανόνων παραγωγής τριπλετών, χρησιμοποιήθηκε το εργαλείο RDF-Gen για την δημιουργία των τελικών αρχείων που εμπεριέχουν τις σχετικές τριπλέτες. Στο τρέχον κεφάλαιο, παρουσιάζονται τα ερωτήματα - queries που σχεδιάστηκαν για την εξαγωγή των επιθυμητών δεδομένων από τις τριπλέτες αυτές, μέσω των εργαλείων της πλατφόρμας Blazegraph.

4.1 Δεδομένα σχετικά με Covid-19 για την πόλη της Αθήνας

Αρχικά, με την ανάθεση των προθεμάτων - prefixes σε αλφαριθμητικά γίνεται εφικτή η συντομογραφία των URIs στις τριπλέτες ή τα SPARQL ερωτήματα. Σκοπός της δημιουργίας του πρώτου ερωτήματος – query είναι η εύρεση των ημερήσιων ενδείξεων που σχετίζονται με τα νέα επιβεβαιωμένα κρούσματα και θανάτους που οφείλονται στον Covid-19.

Τη λέξη κλειδί SELECT ακολουθούν τα ονόματα των μεταβλητών που επιλέχθηκαν για εμφάνιση και εμπλέκονται στις τριπλέτες ενδιαφέροντος. Μέσα στις αγκύλες που ακολουθούν την λέξη κλειδί WHERE, εμπεριέχονται όλες οι σχέσεις - τριπλέτες που πρέπει να ικανοποιούνται για την εξαγωγή της ζητούμενης πληροφορίας.

Η τριπλέτα ?datecity mod:hasNewCases ?NewCases υποδεικνύει όλες τις εγγραφές datecity που σχετίζονται με τιμές NewCases μέσω του κατηγορήματος hasNewCases που ανήκει στην οντολογία mod: <http://www.unipi.gr/ds/msc_bigData/covid19_reports_

and_mob.owl#> που παρουσιάστηκε στην υποενότητα 3.4. Αντίστοιχα, η τριπλέτα ?datecity mod:hasNewDeaths ?NewCases δηλώνει ότι μία εγγραφή datecity συνδέεται μέσω μίας ορισμένης σχέσης hasNewDeaths με μία τιμή NewDeaths που ανήκει στην οντολογία mod:<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_ mob.owl#>. Τέλος, η τριπλέτα ?datecity dul:hasLocation mod:Athens κάνει χρήση μία γενικής σχέσης που προέρχεται από την νέα, επαυξημένη οντολογία και ορίζει την τοποθεσία του αντικειμένου datecity. Σημειώνεται ότι το στιγμιότυπο mod:Athens ορίζεται στην οντολογία που δημιουργήθηκε στα πλαίσια αυτής της διπλωματικής εργασίας, ενώ το

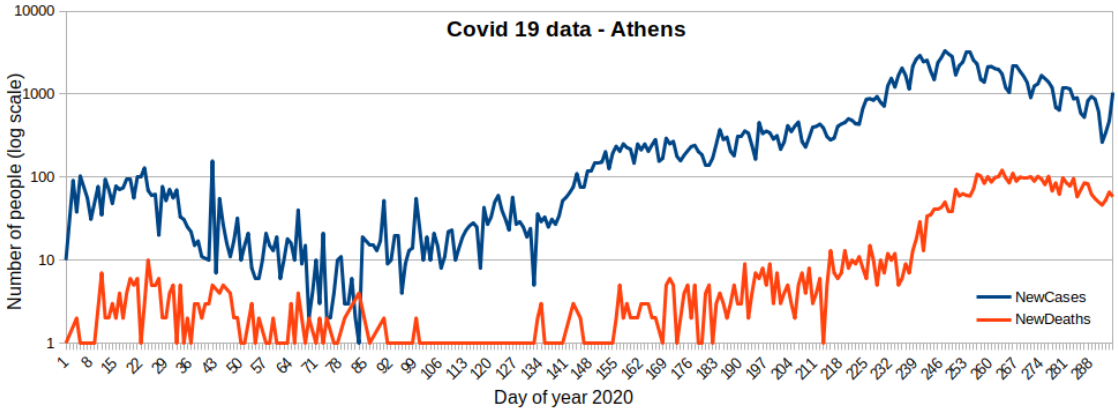
στγμιότυπο dul:hasLocation προέρχεται από την πύλη [30] του Σημασιολογικού Ιστού που σχετίζεται με τα μοτίβα σχεδιασμού οντολογιών. Στην Εικόνα 35 παρουσιάζεται η έξοδος από την εκτέλεση του ερωτήματος σε μορφή πίνακα, ενώ στην Εικόνα 36 τα αποτελέσματα παρουσιάζονται σε μορφή γραφήματος.

Πίνακας 6: Ενδεικτικό ερώτημα – query που επιστρέφει τον αριθμό των νέων κρουσμάτων και θανάτων για την πόλη της Αθήνας.

```
PREFIX dul: <http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#>
PREFIX mod: <http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#>
SELECT  ?datecity ?NewCases ?NewDeaths
WHERE {
?datecity mod:hasNewCases ?NewCases .
?datecity mod:hasNewDeaths ?NewDeaths .
?datecity dul:hasLocation mod:Athens.
}
```

datecity	NewCases	NewDeaths
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#covidReport20200311_Athens>	10	1
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#covidReport20200312_Athens>	0	0
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#covidReport20200316_Athens>	0	0
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#covidReport20200318_Athens>	31	0
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#covidReport20200319_Athens>	0	1
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#covidReport20200406_Athens>	20	6
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#covidReport20200412_Athens>	33	5
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#covidReport20200413_Athens>	31	1

Εικόνα 35: Αποτελέσματα από την εκτέλεση του πρώτου ερωτήματος – query στην πλατφόρμα του εργαλείου Blazegraph.



Εικόνα 36: Αποτελέσματα από την εκτέλεση του πρώτου ερωτήματος - query σε μορφή γραφήματος.

4.2 Μέσος όρος συγκέντρωσης O3 και PM25 ανά πόλη

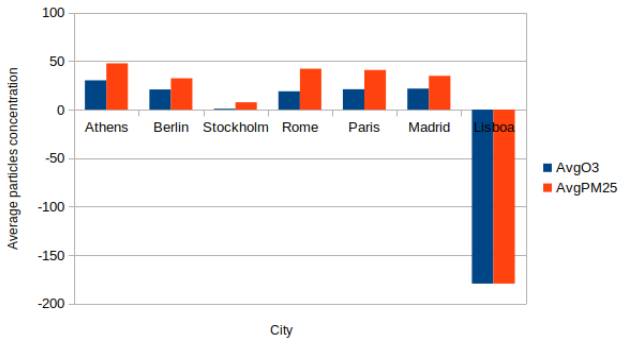
Με το ερώτημα - query που παρουσιάζεται στο Πίνακας 7, υπολογίζεται ο μέσος όρος των τιμών του O3 και του PM25 για κάθε μία από τις πόλεις-πιλότους που χρησιμοποιήθηκαν για τους σκοπούς της παρούσας διπλωματικής εργασίας.

Πίνακας 7: Ενδεικτικό ερώτημα – query που επιστρέφει το μέσο όρο συγκέντρωσης O3 και PM25 για κάθε πόλη – πιλότο της παρούσας διπλωματικής εργασίας.

```
PREFIX mod: <http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#>
SELECT (AVG(?o3) AS ?AvgO3) (AVG(?pm) AS ?AvgPM25) ?loc
WHERE {
?p mod:hasO3 ?o3 .
?p mod:hasPM25 ?pm .
?p dul:hasLocation ?loc
}
GROUP BY ?loc
```

loc	AvgO3	AvgPM25
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#Athens>	30.213598	47.72789
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#Berlin>	20.819256	32.33446
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#Stockholm>	0.99002796	7.601246
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#Rome>	18.941435	42.211838
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#Paris>	21.003744	40.91589
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#Madrid>	21.637995	34.809967
<http://www.unipi.gr/ds/msc_bigData/covid19_reports_and_mob.owl#Lisboa>	-179.43925	-179.43925

Εικόνα 37: Αποτελέσματα από την εκτέλεση του δεύτερου ερωτήματος – query στην πλατφόρμα του εργαλείου Blazegraph.



Εικόνα 38: Παρουσίαση των δεδομένων που προέκυψαν από το δεύτερο ερώτημα – query σε μορφή διαγράμματος στηλών.

4.3 Εύρεση τοποθεσίας νοσοκομείων για την πόλη της Αθήνας

Το ερώτημα-query του Πίνακας 8 χρησιμοποιείται για την εύρεση των συντεταγμένων των νοσοκομείων που υπάρχουν σε μία πόλη όπως η Αθήνα στην συγκεκριμένη περίπτωση. Τα δεδομένα προέρχονται από το Wikidata [29]. Έτσι, στην πρώτη γραμμή του Πίνακας 8 γίνονται οι αντίστοιχες δηλώσεις των prefixes. Στην δεύτερη γραμμή επιλέγονται τα δεδομένα προς εμφάνιση καθώς στην τρίτη περιλαμβάνονται όλες οι απαραίτητες συνθήκες που πρέπει να πληρούνται με την μορφή τριπλετών.

Σημειώνονται τα παρακάτω:

- Wd: Q16917 αντιπροσωπεύει την οντότητα των νοσοκομείων,
- Wd: Q41 είναι μοναδικό αναγνωριστικό για την χώρα της Ελλάδας,
- Wd: Q1524 αποτελεί μοναδικό αναγνωριστικό για την πόλη της Αθήνας,
- Τα υπόλοιπα γνωρίσματα που χρησιμοποιούν το γράμμα P αναφέρονται στις διάφορες σχέσεις της οντολογίας στην οποία ανήκουν

Πίνακας 8: Ενδεικτικό ερώτημα - query που επιστρέφει την τοποθεσία των νοσοκομείων στη περιοχή της Αθήνας

```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
SELECT DISTINCT ?item ?geo ?Country
WHERE {
  ?item (wdt:P31/(wdt:P279*)) wd:Q16917;
    wdt:P625 ?geo.
  ?Countrywdt:P17 wd:Q41;
    wdt:P625 ?geo.
  ?City wdt:P131 wd:Q1524;
    wdt:P625 ?geo.
}
```

Στην Εικόνα 39 παρουσιάζονται τα αποτελέσματα του τρίτου ερωτήματος σε μορφή πίνακα, του οποίου η πρώτη στήλη δείχνει το μοναδικό αναγνωριστικό του νοσοκομείου, η δεύτερη την γεωγραφική του θέση και η τρίτη το μοναδικό αναγνωριστικό της αντίστοιχης χώρας. Στην Εικόνα 40 παρουσιάζεται η τοποθεσία των νοσοκομείων πάνω στον χάρτη με την μορφή κόκκινων σημείων.

item	geo	Country
wd:Q21626578	Point(23.755355 37.984826)	wd:Q21626578
wd:Q21626578	Point(23.7531612 37.9848286)	wd:Q21626578
wd:Q15762293	Point(23.787519 38.029543)	wd:Q15762293
wd:Q21626593	Point(23.875058 38.051006)	wd:Q21626593
wd:Q21626577	Point(23.765596 37.98345)	wd:Q21626577
wd:Q30260516	Point(23.758 37.982608)	wd:Q30260516

Εικόνα 39: Αποτελέσματα από την εκτέλεση του τρίτου ερωτήματος - query στην πλατφόρμα του εργαλείου Blazegraph.

Εικόνα 40: Τοποθεσία των νοσοκομείων της Αθήνας σε μορφή χάρτη.

4.4 Εύρεση τοποθεσίας αθλητικών κέντρων στην περιοχή της Αθήνας

Ο Πίνακας 9 παρουσιάζει τη δομή του ερωτήματος – query που δημιουργήθηκε για την εύρεση της τοποθεσίας των αθλητικών κέντρων (γηπέδων) στην περιοχή της Αθήνας. Όπως φαίνεται στην πρώτη γραμμή του πίνακα, τα σχετικά δεδομένα προέρχονται από

Πίνακας 9: Ενδεικτικό ερώτημα - query που επιστρέφει την τοποθεσία αθλητικών εγκαταστάσεων στην περιοχή της Αθήνας.

```
PREFIX osmt: <https://wiki.openstreetmap.org/wiki/Key:>
PREFIX osmm: <https://www.openstreetmap.org/meta/>
SELECT *
WHERE {
  ?pitch osmt:leisure "pitch" .
  SERVICE wikibase:around {
    ?pitch osmm:loc ?coordinates.
    bd:serviceParam wikibase:center "Point(23.7275 37.9838)"^^geo:wktLiteral.
    bd:serviceParam wikibase:radius "10". # kilometers
    bd:serviceParam wikibase:distance ?distance.
  }
}
```

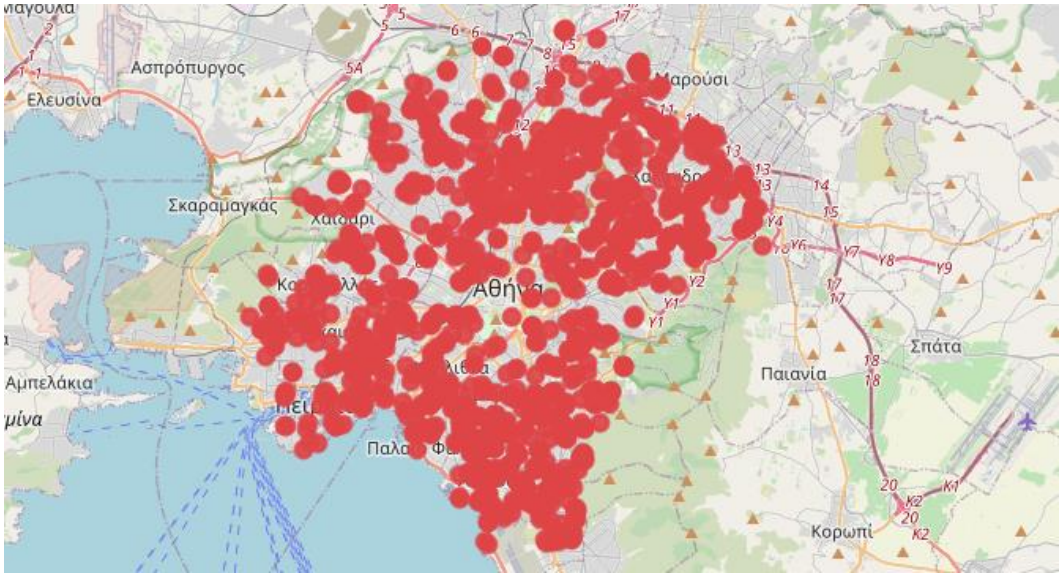
την πηγή [29].

Η τριπλέτα ?pitch osmt:leisure "pitch" ορίζει ότι αναζητείται πληροφορία που σχετίζεται με δραστηριότητες ελεύθερου χρόνου και έχουν την ετικέτα pitch. Επιπρόσθετα, η ετικέτα ?pitch osmm:loc ?coordinates ορίζει την τοποθεσία των ευρημάτων που επιτυγχάνεται με την χρήση του SERVICE που υπολογίζει την απόσταση χρησιμοποιώντας ως παραμέτρους το κέντρο της αναζήτησης και την επιθυμητή ακτίνα γύρω από αυτό.

Στην Εικόνα 41 παρουσιάζονται μερικά από τα αποτελέσματα του ερωτήματος σε μορφή πίνακα, ενώ στην Εικόνα 42 παρουσιάζονται οι ακριβείς τοποθεσίες πάνω στον χάρτη με μορφή κόκκινων σημείων.

pitch	coordinates	distance
osmway:211110371 edit	Point(23.6208938 37.9684882)	9.498
osmway:211110372 edit	Point(23.6209816 37.9685023)	9.49
osmway:212874434 edit	Point(23.7183887 37.8960386)	9.791
osmway:601561477 edit	Point(23.719794 37.8949466)	9.903
osmway:403378906 edit	Point(23.616104 37.9730475)	9.836
osmway:705081322 edit	Point(23.7192571 37.9026585)	9.051
osmway:705081323 edit	Point(23.719328 37.9026565)	9.051

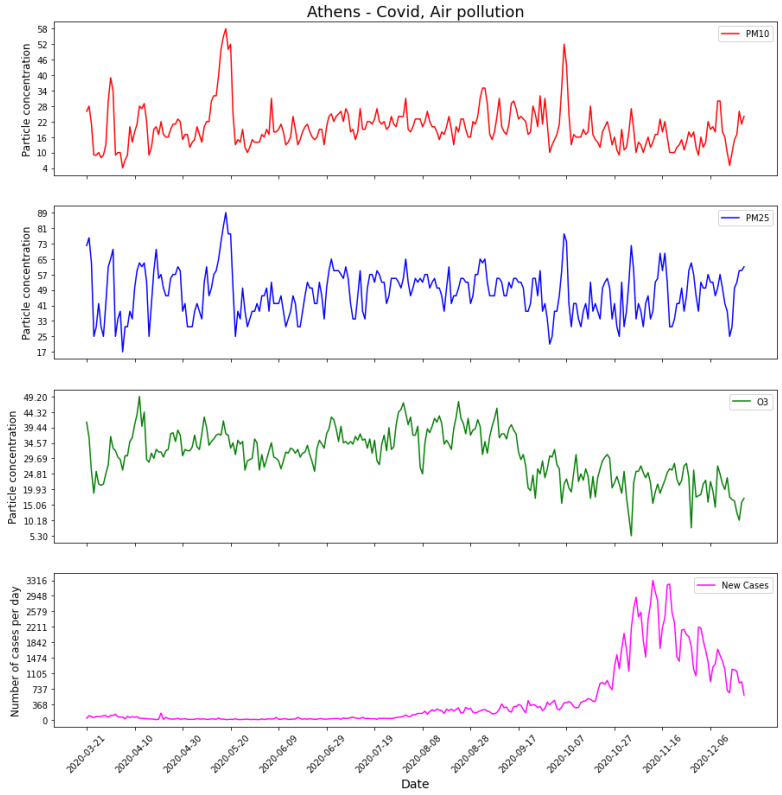
Εικόνα 41: Αποτελέσματα από την εκτέλεση του τέταρτου ερωτήματος - query στην πλατφόρμα του εργαλείου Blazegraph.



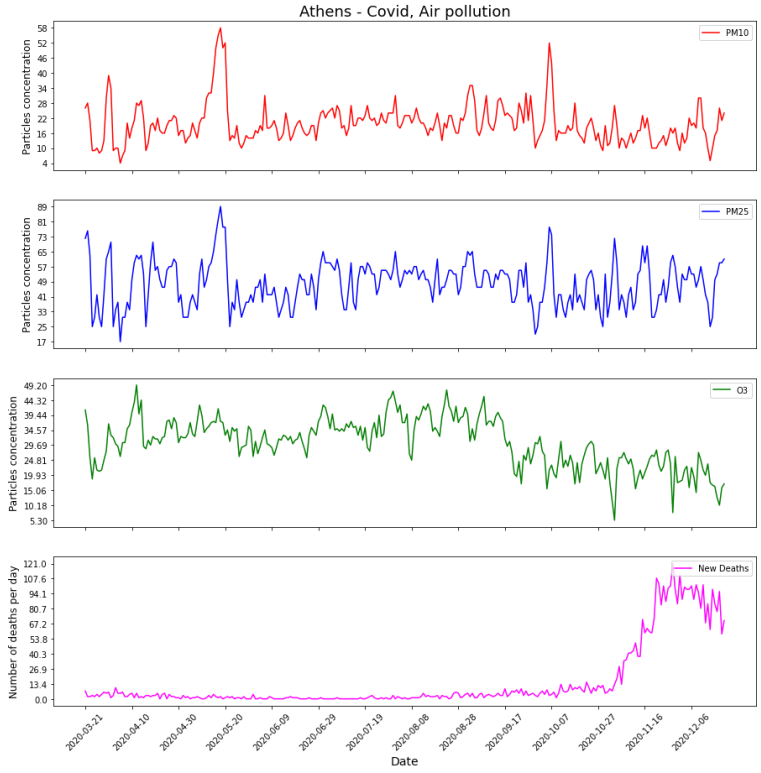
Εικόνα 42: Παρουσίαση αθλητικών κέντρων στην περιοχής της Αθήνας σε μορφή χάρτη

4.5 Παρατηρήσεις

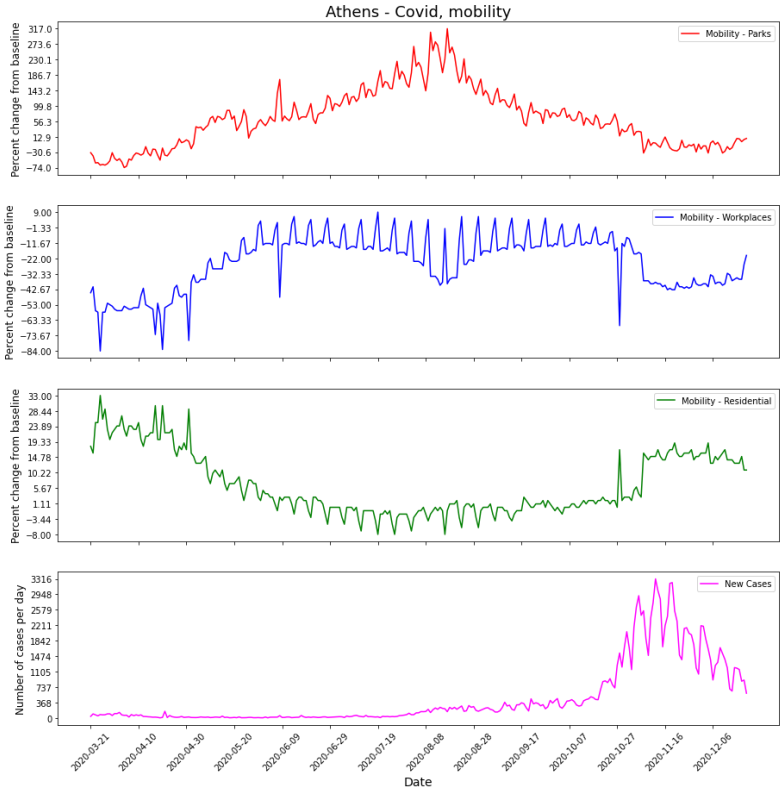
Ο συνδυασμός των βασικών λειτουργιών που υλοποιήθηκαν με τα SPARQL queries και παρουσιάστηκαν παραπάνω καθιστά εφικτή την δημιουργία πιο σύνθετων αποτελεσμάτων που συνεισφέρουν στην αντιμετώπιση των ερωτημάτων που τέθηκαν στην υπο-ενότητα 3.1.1. Η Εικόνα 43 παραθέτει τη γραφική παράσταση των νέων ημερήσιων κρουσμάτων παράλληλα με ατμοσφαιρικά δεδομένα για το έτος 2020 και την περιοχή της Αθήνας. Αντίστοιχα, η Εικόνα 44 δείχνει τους ημερήσιους θανάτους από Covid-19 σε συνδυασμό με τα ίδια ατμοσφαιρικά δεδομένα. Παρόλο που μία άμεση σύνδεση δεν είναι εμφανής με την πρώτη εικόνα, ένα εξειδικευμένος επιστήμονας-ερευνητής θα επωφεληθεί από την παροχή μίας τέτοιας πληροφορίας και ίσως καταλήξει σε μια εμπεριστατωμένη άποψη. Από την άλλη πλευρά, η Εικόνα 45 και η Εικόνα 46 φαίνεται να προσφέρουν πιο ξεκάθαρες ενδείξεις. Για παράδειγμα, κατά το τελευταίο διάστημα του έτους παρατηρείται σημαντική αύξηση τόσο στα ημερήσια κρούσματα όσο και στους ημερήσιους θανάτους. Την ίδια περίοδο παρουσιάζεται αισθητή μείωση στις μετακινήσεις τον ανθρώπων προς τους χώρους εργασίες και τα πάρκα, ενώ αντίθετα αυξάνεται η παραμονή των ανθρώπων στους χώρους κατοικίας.



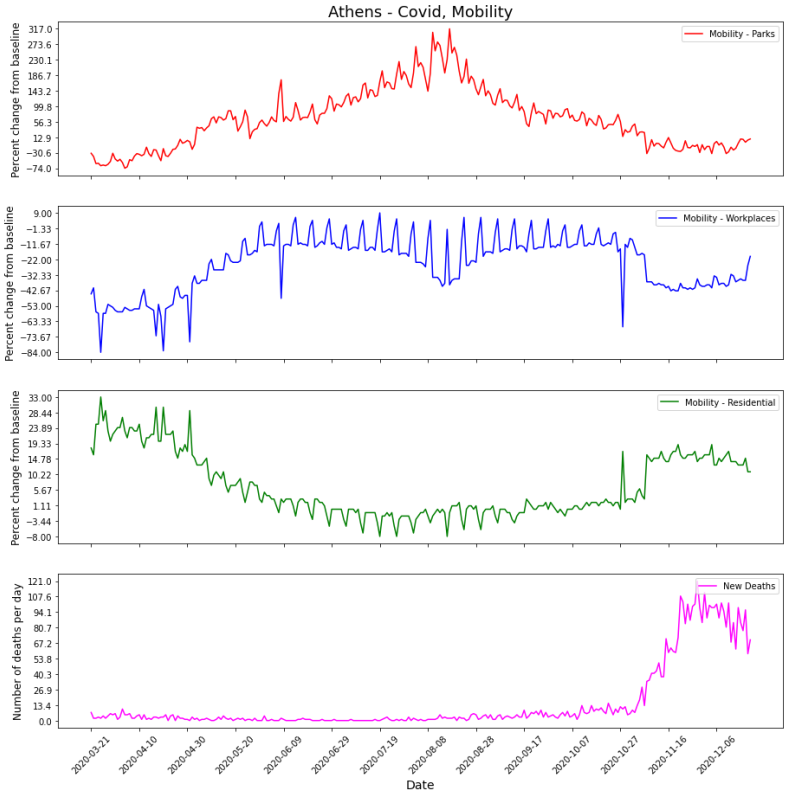
Εικόνα 43: Παράθεση νέων ημερησίων κρουσμάτων Covid-19 και περιβαλλοντικών δεδομένω για την περιοχή της Αθήνας



Εικόνα 44: Παράθεση νέων ημερησίων θανάτων λόγω Covid-19 και περιβαλλοντικών δεδομένω για την περιοχή της Αθήνας



Εικόνα 45: Παράθεση νέων ημερησίων κρουσμάτων Covid-19 και δεδομένων μετακίνησης για την περιοχή της Αθήνας



Εικόνα 46: Παράθεση νέων ημερησίων θανάτων από Covid-19 και δεδομένων μετακίνησης για την περιοχή της Αθήνας

5 Συμπεράσματα – Μελλοντικές επεκτάσεις

Στην παρούσα εργασία συλλέχθηκαν δεδομένα από επίσημες πηγές σχετικά με την πανδημία Covid-19, ατμοσφαιρικής μόλυνσης και μετακίνησης. Από αυτά δημιουργήθηκε η αντίστοιχη οντολογία, κάνοντας χρήση της γλώσσας OWL και του εργαλείου Protégé Στην συνέχεια, με την χρήση κατάλληλων λεξικών και γραμματικής δημιουργήθηκαν οι επιθυμητές RDF τριπλέτες ενσωματωμένες στα αντίστοιχα αρχεία, τα οποία μεταφορτώθηκαν στο αποθετήριο τριπλετών Blazegraph. Το επόμενο βήμα σχετίζεται με τον σχεδιασμό και την κατασκευή ερωτημάτων – queries για την εξαγωγή των επιθυμητών αποτελεσμάτων από της τριπλέτες. Επιπρόσθετα, με χρήση SPARQL ερωτημάτων προς εξωτερικές πηγές, ανακτήθηκαν επιπλέον δεδομένα από συγκεκριμένα σημεία του σημασιολογικού ιστού, τα οποία σχετίζονται με τον αριθμό νοσοκομείων και χώρων αθλητισμού και πυκνότητα πληθυσμού για τις πόλεις-πιλότους της παρούσας διπλωματικής. Ο συνδυασμός όλων των αρχικά ετερογενών δεδομένων οδήγησε στην εξαγωγή των αποτελεσμάτων που παρουσιάζονται στην προηγούμενη ενότητα. Πρέπει να σημειωθεί ότι στα πλαίσια της παρούσας εργασίας δημιουργήθηκε μία ολοκληρωμένη διαδικασία μετατροπής των αρχικών δεδομένων σε μία ομογενή οντολογία που επιδέχεται ερωτήματα – queries σε μορφή SPARQL.

Μέσω αυτής της διπλωματικής εργασίας πραγματοποιήθηκε μία στοχευμένη συλλογή δεδομένων, η επεξεργασία τους, η δόμησή τους με πρότυπα του σημασιολογικού ιστού και η εξαγωγή χρήσιμης πληροφορίας μέσω της ανάλυσής τους. Τα ευρήματα αυτής της διπλωματικής εργασίας μπορούν να χρησιμοποιηθούν μελλοντικά για τη διάθεσή των σχετικών πληροφοριών στο σημασιολογικό ιστό. Παράλληλα, μία μελλοντική διερεύνηση θα μπορούσε να περιλαμβάνει αλγορίθμους μηχανική μάθησης, προκειμένου να εξαχθούν πιο ακριβή συμπεράσματα και για πιο περίπλοκους συνδυασμούς.

Βιβλιογραφία

[1]. Collins, Allan M., and M. Ross Quillian. "Retrieval time from semantic memory." Journal of verbal learning and verbal behavior 8.2 (1969): 240-247.

[2]. Collins, Allan M., and Elizabeth F. Loftus. "A spreading-activation theory of semantic processing." *Psychological review* 82.6 (1975): 407.

[3]. Quillian, M. Ross. "Word concepts: A theory and simulation of some basic semantic capabilities." *Behavioral science* 12.5 (1967): 410-430.

[4]. Minsky, Marvin. "Semantic information processing." (1982).

[5]. Bhiri, Sami, et al. "Semantic web services for satisfying SOA requirements." Advances in Web Semantics I. Springer, Berlin, Heidelberg, 2008. 374-395.

[6]. Greenberg, Jane, Stuart Sutton, and D. Grant Campbell. "Metadata: a fundamental component of the semantic web." Bulletin of the American Society for Information Science and Technology 29.4 (2003): 16-16.

[7]. Simperl, Elena. "Reusing ontologies on the Semantic Web: A feasibility study." Data & Knowledge Engineering 68.10 (2009): 905-925.

[8]. Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web." *Scientific american* 284.5 (2001): 34-43.

[9]. "Semantic Web - XML2000, slide 10". W3C. Retrieved 2008-05-13.

[10]. Berners-Lee, Tim J. "The world-wide web." Computer networks and ISDN systems 25.4-5 (1992): 454-459.

[11]. Antoniou, Grigoris, and Frank Van Harmelen. A semantic web primer. MIT press, 2004.

[12]. Weibel, Stuart, et al. "Dublin core metadata for resource discovery." Internet Engineering Task Force RFC 2413.222 (1998): 132.

[13]. Brickley, Dan, Ramanathan V. Guha, and Andrew Layman. "Resource description framework (RDF) schema specification." (1999).

[14]. Musen, M.A. The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.

[15]. Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 22 July 2004. Web. 10 Aug. 2004.

[16]. Auer, Sören, et al. "Dbpedia: A nucleus for a web of open data." The semantic web. Springer, Berlin, Heidelberg, 2007. 722-735.

[17]. Pérez, Jorge, Marcelo Arenas, and Claudio Gutierrez. "Semantics and complexity of SPARQL." ACM Transactions on Database Systems (TODS) 34.3 (2009): 1-45.

[18]. McGuinness, Deborah L., and Frank Van Harmelen. "OWL web ontology language overview." W3C recommendation 10.10 (2004): 2004.

[19]. Beckett, Dave, and Brian McBride. "RDF/XML syntax specification (revised)." W3C recommendation 10.2.3 (2004).

[20]. https://www.w3.org/Consortium/

[21]. Coronavirus 2019-nCoV, CSSE . Coronavirus 2019-nCoV Global Cases by Johns Hopkins CSSE.

[22]. https://ourworldindata.org/

[23]. https://www.google.com/covid19/mobility/

[24]. https://waqi.info/

[25]. https://blazegraph.com/

[26]. Seber, George AF, and Alan J. Lee. Linear regression analysis. John Wiley & Sons, 2012.

[27]. "Main Page." OpenStreetMap Wiki, . 19 Jul 2020, 09:05 UTC. 20 Feb 2022, 16:40 https://wiki.openstreetmap.org/w/index.php?title=Main_Page&oldid=2013332 .

[28]. Santipantakis, Georgios M., et al. "Rdf-gen: Generating RDF from streaming and archival data." Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. 2018.

[29]. https://www.wikidata.org/wiki/Wikidata:Main_Page

[30]. http://www.ontologydesignpatterns.org/