ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΔΙΟΙΚΗΣΗΣ

Διπλωματική Εργασία

ΑΝΑΛΥΤΙΚΗ ΑΝΟΙΧΤΩΝ ΣΥΝΔΕΔΕΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ

του

ΓΕΡΑΣΙΜΟΥ ΑΝΤΩΝΙΟΥ ΤΟΥ ΛΕΩΝΙΔΑ

Υποβλήθηκε ως προαπαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος ειδίκευσης στα Πληροφοριακά Συστήματα

Εποπτεύοντες καθηγητές:
Κωνσταντίνος Ταραμπάνης
Καθηγητής
kat@uom.edu.gr
http://islab.uom.gr

Ευθύμιος Ταμπούρης
Καθηγητής
tambouris@uom.edu.gr
http://islab.uom.gr

Οκτώβριος 2018

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ευκαιρία που μου δόθηκε μέσω αυτής της διπλωματικής, θα ήθελα να εκφράσω τις ευχαριστίες μου πρωτίστως στους καθηγητές μου Κωνσταντίνο Ταραμπάνη και Ευθύμιο Ταμπούρη καθώς και σε όλους τους καθηγητές του ΔΠΜΣ για όλο τον χρόνο και τις γνώσεις που μου διέθεσαν. Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στους γονείς μου για την υποστήριξή τους καθώς και την σύντροφό μου Δήμητρα για την υπομονή και την βοήθειά της.

ΠΕΡΙΛΗΨΗ

Η εποχή μας έχει χαρακτηριστεί ως η εποχή της πληροφορίας. Η αυτοματοποιημένη συλλογή δεδομένων μέσω αισθητήρων, συστημάτων παρακολούθησης και smartphone εφαρμογών οδηγεί σταδιακά στο λεγόμενο «datafication». Στο πλαίσιο αυτό αναπτύχθηκε η προσέγγιση των ανοιχτών συνδεδεμένων δεδομένων του Σημασιολογικού Ιστού, η ανάλυση και εκμετάλλευση των οποίων βοηθάει δυνητικά τους χρήστες στο να αποκτήσουν διορατικότητα. Στην εργασία μας θα ασχοληθούμε με τα ανοιχτά συνδεδεμένα δεδομένα δίνοντας έμφαση στα ανοιχτά συνδεδεμένα νομικά δεδομένα και θα μελετήσουμε μεθόδους εξόρυξης και ανάλυσης των δεδομένων με στόχο να συσχετίσουμε τις πληροφορίες χρησιμοποιώντας τα κατάλληλα εργαλεία ώστε να εξαχθεί γνώση.

Αποτέλεσμα της μελέτης μας ήταν η εξοικείωση με βασικές έννοιες, πρότυπα και τεχνικές των ανοιχτών συνδεδεμένων δεδομένων, η εμβάθυνση στις οντολογίες των Φιλανδικών ανοιχτών συνδεδεμένων νομικών δεδομένων, η πρόσβαση και εξόρυξη των δεδομένων μέσω SPARQL ερωτημάτων που συντάξαμε κατά την συγγραφή της παρούσας και η ανάλυση των αποτελεσμάτων με την επέκταση LOD extension του RapidMiner.

Η εξεύρεση σωστά δομημένων και ποιοτικών συνόλων δεδομένων αποδείχθηκε δύσκολο έργο, γεγονός που μας οδηγεί στο συμπέρασμα ότι οι αρχές και τα πρότυπα του Σημασιολογικού Ιστού (Web 3.0) δεν εφαρμόστηκαν σωστά μέχρι σήμερα και έτσι από τα δισεκατομμύρια των δεδομένων που παράγονται καθημερινά μόνο ένα μικρό ποσοστό γίνεται αντικείμενο περαιτέρω μελέτης και ανάλυσης και μετουσιώνεται σε γνώση. Καθώς λοιπόν βαδίζουμε προς τον Συμβιωτικό Ιστό (Web 4.0) πρέπει να δοθεί έμφαση στη σωστή αποθήκευση, δημοσίευση και επωφελή εκμετάλλευση των δεδομένων.

ABSTRACT

Our era has been characterized as the era of information. The automated data collection through sensors, monitoring systems and smartphone applications progressively leads to the so-called «datafication». In this context, the approach of linked open data of the Semantic Web was developed, the analysis and exploitation of which potentially helps users to gain insight. In our essay we will deal with linked open data, focusing on linked open legal data and we will study methods of mining and analyzing data in order to relate the information using the appropriate tools to extract knowledge.

The result of our study was the familiarization with basic concepts, patterns and techniques of linked open data, the deepening on the ontologies of the Finnish linked open legal data, the access and extraction of data through SPARQL queries we drafted during the drafting of this essay, and the analysis of the results with the RapidMiner LOD extension tool.

Finding properly structured and qualitative datasets proved to be a difficult task, leading us to the conclusion that the principles and standards of the Semantic Web (Web 3.0) have not been properly implemented so far, and so out of the billions of data produced on a daily basis only a small percentage is subject to further study and analysis and is transformed into knowledge. Therefore, as we move towards the Symbiotic Web (Web 4.0), we must emphasize on the proper storage, publication, and profitable exploitation of the data.

Περιεχόμενα

Κατάλογος Εικόνων:

Κατάλογος Πινάκων:

1. ΕΙΣΑΓΩΓΗ

1.1 Περιγραφή του προβλήματος

Η ταχεία πρόοδος στην τεχνολογία των πληροφοριών του 21ου αιώνα συνδέεται άμεσα με την ακραία αύξηση της συλλογής και αποθήκευσης ψηφιακών δεδομένων. Αυτό είναι το αποτέλεσμα της δικτύωσης και της παγκοσμιοποίησης, της συνεχούς βελτίωσης της πληροφορικής, των εξαιρετικά εξελιγμένων βάσεων δεδομένων, του Διαδικτύου ως πλατφόρμας επικοινωνίας και της τεράστιας επέκτασης της αυτοματοποιημένης συλλογής δεδομένων μέσω αισθητήρων, συστημάτων παρακολούθησης και εφαρμογών κινητών και smartphone που οδηγεί σταδιακά στο λεγόμενο **«datafication»** και στην **«υπερφόρτωση πληροφοριών»** [*Lausch, Schmidt & Tischendorf,* 2015]. Σύμφωνα με την έκτη ετήσια έκθεση της εταιρείας παροχής λογισμικού – υπηρεσιών διαδικτύου DOMO [https://www.domo.com] που δημοσιεύθηκε τον Ιούνιο του 2018 περίπου το 90% των δεδομένων στον κόσμο δημιουργήθηκε κατά τη διετία 2015-2017. Πάνω από 2,5 quintillion bytes (2,500,000,000,000,000,000) δεδομένων δημιουργούνται κάθε μέρα με διαρκή αυξητική τάση, ενώ μέχρι το 2020 εκτιμάται ότι 1,7MB δεδομένων θα δημιουργούνται κάθε δευτερόλεπτο για κάθε άτομο στη γη [https://www.domo.com/news/press/domo-releases-sixth-annual-data-never-sleeps-infographic].

Τα δεδομένα όμως από μόνα τους δεν μπορούν να γίνουν εύκολα κατανοητά, ούτε να αξιοποιηθούν κατάλληλα από τους χρήστες, γεγονός που οδήγησε τον εφευρέτη του Παγκόσμιου Ιστού (World Wide Web) και διευθυντή του W3C (World Wide Web Consortium) Tim Berners-Lee στη σύλληψη της ιδέας ενός ιστού δεδομένων (web of data) ή Σημασιολογικού Ιστού (Semantic Web) όπου τα δεδομένα είναι δημοσιευμένα σύμφωνα με συγκεκριμένες πρακτικές ώστε να μπορούν να συνδεθούν μεταξύ τους σε δίκτυο κατά τρόπο ώστε ένα άτομο ή ένας υπολογιστής να μπορεί να τα εξερευνήσει, να τα εκμεταλλευτεί και να βρει περαιτέρω συσχετισμένες πληροφορίες [*Berners-Lee,* 2009]. Το μεγάλο πλεονέκτημα του Σημασιολογικού Ιστού εντοπίζεται στην ενοποίηση των δεδομένων από διαφορετικές πηγές ώστε και οι υπολογιστές να κατανοούν καλύτερα τις πληροφορίες που βρίσκονται στον Ιστό αλλά και οι χρήστες να έχουν ένα πολύτιμο εργαλείο εύρεσης, ένωσης και επεξεργασίας των πληροφοριών.

Η προσέγγιση των ανοικτών συνδεδεμένων δεδομένων του Σημασιολογικού Ιστού ευνοεί παράλληλα την ανάπτυξη νέων μεθόδων και νέων δυνατοτήτων διεπιστημονικής ανάλυσης, μοντελοποίησης και εκμετάλλευσης των δεδομένων με την χρήση

διαδραστικών εργαλείων εξόρυξης δεδομένων που μπορούν να χρησιμοποιηθούν από χρήστες με ελάχιστες γνώσεις προγραμματισμού για τη διενέργεια ερευνών και την εξαγωγή χρήσιμων συμπερασμάτων.

Παρά την παραγωγή και διαθεσιμότητα τεράστιων ποσοτήτων δεδομένων, τις ταχύτατες τεχνολογικές εξελίξεις και τη δημιουργία νέων τεχνικών εξόρυξης και ανάλυσης, η εκμετάλλευση των Ανοιχτών Συνδεδεμένων Δεδομένων είναι ακόμη σε πρώιμο στάδιο. Αυτό οφείλεται κυρίως στα τεχνικά προβλήματα των σημείων στον Ιστό που παρέχουν πρόσβαση στα δεδομένα μέσω ερωτημάτων (πχ SPARQL endpoints κλπ) τα οποία σε αρκετές περιπτώσεις δεν λειτουργούν ή έχουν καταργηθεί. Ένα άλλο πρόβλημα είναι ότι υπάρχουν σετ Ανοιχτών Συνδεδεμένων Δεδομένων που δεν είναι επικαιροποιημένα ή είναι ελλιπή. Συνεπώς, όλες αυτές οι πληροφορίες δεν μπορούν να αξιοποιηθούν από τους χρήστες και να τους βοηθήσουν στην λήψη ορθών επιχειρησιακών και στρατηγικών αποφάσεων, στο να κάνουν προβλέψεις και γενικότερα στο να αποκτήσουν διορατικότητα (insight).

1.2 Αντικείμενο και στόχοι της μελέτης

Στο πλαίσιο της παρούσας εργασίας θα εξετάσουμε τη δομή των ανοιχτών συνδεδεμένων δεδομένων και θα αναφερθούμε ειδικότερα στις πρωτοβουλίες που έχουν αναπτυχθεί τα τελευταία χρόνια σχετικά με την ελεύθερη πρόσβαση και διασύνδεση των νομικών δεδομένων. Ο βασικός στόχος της μελέτης είναι να παρουσιάσουμε πώς τα ανοιχτά συνδεδεμένα δεδομένα μπορούν να αξιοποιηθούν στην εξόρυξη δεδομένων (data mining), να συσχετιστούν μεταξύ τους και να αναλυθούν με τις κατάλληλες μεθόδους ώστε από ένα απλό σύνολο δεδομένων να εξαχθούν επιπλέον πληροφορίες και γνώση.

Η εργασία αυτή θα συνεισφέρει ιδίως στην εξοικείωση του αναγνώστη με βασικές αρχές, ορολογίες και πρότυπα δημοσίευσης και διασύνδεσης των ανοιχτών συνδεδεμένων δεδομένων και στην κατανόηση της διαλειτουργικότητας και χρησιμότητάς τους. Θα προσφέρει επίσης ενδιαφέρουσες πληροφορίες σχετικά με τις προσπάθειες των διαφόρων κυβερνήσεων να εφαρμόσουν αυτές τις αρχές και τεχνικές στα νομικά δεδομένα που βρίσκονται ακόμη σε αρχικό στάδιο σε αντίθεση με άλλες κατηγορίες κυβερνητικών δεδομένων.

Περαιτέρω, θα περιγραφεί ο τρόπος συλλογής και αξιολόγησης των δεδομένων με την χρήση του εργαλείου RapidMiner και ιδίως της επέκτασης LOD για την εξαγωγή χρήσιμων συμπερασμάτων. Επιμέρους στόχος είναι να παρουσιάσουμε τις δυσχέρειες

που συναντήσαμε κατά την προσπάθεια άντλησης και ανάλυσης σωστά δομημένων και ποιοτικών συνόλων δεδομένων και τις διαπιστώσεις μας σχετικά με το επίπεδο υλοποίησης της ιδέας των ανοιχτών συνδεδεμένων δεδομένων.

1.3 Περιεχόμενα της μελέτης

Η παρούσα εργασία αποτελείται από πέντε κεφάλαια που διαιρούνται σε ενότητες και υποενότητες για τη διευκόλυνση του αναγνώστη και σύμφωνα με τη ροή της πραγματοποιηθείσας έρευνας. Ειδικότερα, το δεύτερο κεφάλαιο περιλαμβάνει την ανασκόπηση της βιβλιογραφίας, τον τρόπο συλλογής των πηγών και τα ευρήματα της επισταμένης μελέτης μας αναφορικά με τα ανοιχτά συνδεδεμένα δεδομένα και ιδίως τα νομικά ανοιχτά συνδεδεμένα δεδομένα, την επιστήμη της εξόρυξης δεδομένων και το εργαλείο του RapidMiner. Η συγγραφή του κεφαλαίου της βιβλιογραφικής ανασκόπησης ακολουθεί την σειρά της έρευνας που διεξήχθη στις διαθέσιμες πηγές για την εκπόνηση της παρούσας εργασίας. Η συνολική βιβλιογραφία παρατίθεται στο τέλος της μελέτης κατ' αλφαβητική σειρά. Στο αμέσως επόμενο κεφάλαιο, θα παρουσιάσουμε τη μεθοδολογία που χρησιμοποιήθηκε για το θέμα της παρούσας εργασίας, με παράλληλη αναφορά στην συγκεκριμένη τεχνική έρευνας που επιλέξαμε καθώς και στα σύνολα δεδομένων από όπου αντλήσαμε πληροφορίες προς ανάλυση και εξαγωγή συμπερασμάτων. Στο τέταρτο κεφάλαιο θα αναφερθούμε στις πηγές από τις οποίες επιλέξαμε να λάβουμε τα προς ανάλυση δεδομένα και θα παρουσιάσουμε τα ερωτήματα (queries) που συντάξαμε για να αντλήσουμε συγκεκριμένες πληροφορίες από ένα σύνολο δεδομένων. Επίσης, θα παρουσιάσουμε τα ερωτήματα που συντάξαμε για την λήψη συγκεκριμένων μεταβλητών των δεδομένων, θα χρησιμοποιήσουμε το εργαλείο του RapidMiner για την σύνδεση των αποτελεσμάτων με άλλα σύνολα ανοιχτών δεδομένων και θα αναλύσουμε τα πορίσματα της έρευνας. Το πέμπτο και τελευταίο κεφάλαιο περιέχει τα συμπεράσματα της παρούσας εργασίας και διάφορες προτάσεις για μελλοντική έρευνα.

2. ΑΝΑΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

2.1 Εισαγωγή

Η έρευνα μας εκκίνησε με σκοπό να κατανοήσουμε καταρχήν τι είναι τα ανοιχτά συνδεδεμένα δεδομένα, πώς προέκυψαν στην πορεία της εξέλιξης του Διαδικτύου και ποιες είναι οι τεχνικές που πρέπει να ακολουθηθούν προκειμένου να μορφοποιηθούν με κατάλληλο τρόπο στον Ιστό ώστε να γίνει δυνατή η διασύνδεσή τους. Σημαντική βοήθεια για τα παραπάνω παρείχε ο ιστότοπος του W3C [https://www.w3.org], οι διάφορες δημοσιεύσεις και άρθρα του Tim Berners-Lee και φυσικά οι οικείες παραπομπές σε βιβλία και άρθρα. Μεγάλη ήταν και η συνεισφορά της βάσης δεδομένων της βιβλιοθήκης του Πανεπιστημίου Μακεδονίας, όπου έγινε αναζήτηση βιβλιογραφίας με βάση λέξεις - κλειδιά [https://www.lib.uom.gr/dbases/greek/index.php] και φυσικά ο μελετητής Google Scholar.

Ακολούθως, λαμβάνοντας ως βάση για την περαιτέρω έρευνα μας πρόσφατο άρθρο του Enrico Francesconi (2018) σχετικά με το μέλλον της δημοσίευσης των νομικών δεδομένων στον Σημασιολογικό Ιστό, η βιβλιογραφική μας έρευνα κατευθύνθηκε προς την αναζήτηση πληροφοριών σχετικά με τα πλεονεκτήματα διασύνδεσης των νομικών δεδομένων και τις πρωτοβουλίες διαφόρων χωρών προς υλοποίηση της προσέγγισης των Ανοιχτών Συνδεδεμένων Δεδομένων και στα νομικά δεδομένα. Από τις πηγές που συλλέχθηκαν προέκυψε ότι ιδίως οι χώρες της Ευρωπαϊκής Ένωσης έχουν αντιληφθεί την χρησιμότητα της ελεύθερης πρόσβασης και συσχέτισης των νομικών δεδομένων και έχουν προχωρήσει σε εκτέλεση προγραμμάτων, τα σημαντικότερα των οποίων θα παρουσιάσουμε στην αντίστοιχη ενότητα του παρόντος κεφαλαίου.

Τέλος, κατόπιν της κατανόησης των χαρακτηριστικών των ανοιχτών συνδεδεμένων δεδομένων και της στοχευμένης έρευνας στα ανοιχτά συνδεδεμένα νομικά δεδομένα, η βιβλιογραφική μας αναζήτηση επικεντρώθηκε στην μελέτη των διαφόρων μεθόδων εξόρυξης δεδομένων, στη διαδικασία που ακολουθείται για την άντληση των δεδομένων και στην εκμάθηση του εργαλείου RapidMiner ούτως ώστε να γίνει η ανάλυση των δεδομένων που παρατίθεται στο τέταρτο κεφάλαιο της εργασίας.

2.2 ΤΑ ΑΝΟΙΧΤΑ ΣΥΝΔΕΔΕΜΕΝΑ ΔΕΔΟΜΕΝΑ (LINKED OPEN DATA)

2.2.1 Έννοια και χαρακτηριστικά γνωρίσματα των συνδεδεμένων δεδομένων και των ανοιχτών δεδομένων

Ο όρος Συνδεδεμένα Δεδομένα (Linked Data) αναφέρεται σε *«δεδομένα που είναι δημοσιευμένα στον Παγκόσμιο Ιστό κατά τρόπο που είναι αναγνώσιμα από μηχανές, το νόημά τους είναι σαφώς καθορισμένο, συνδέονται με άλλα εξωτερικά σύνολα δεδομένων και είναι με τη σειρά τους προσβάσιμα από εξωτερικά σύνολα δεδομένων»* [*Bizer, Heath, & Berners-Lee*, 2009]. Με τα συνδεδεμένα δεδομένα λοιπόν δημιουργούνται σύνδεσμοι ανάμεσα σε πληροφορίες από διαφορετικές πηγές. Η συντακτική και σημασιολογική διαλειτουργικότητα (semantic interoperability), που συνήθως ορίζεται ως η ικανότητα δύο ή περισσοτέρων εφαρμογών να κατανοούν τα δεδομένα μεταξύ τους, αποτελεί το στόχο των Συνδεδεμένων Δεδομένων στον κόσμο του διαδικτύου.

Η υλοποίηση της σύνδεσης μεταξύ των δεδομένων όμως προϋποθέτει την ύπαρξη και πρόσβαση σε «ανοιχτά δεδομένα» (open data). Σύμφωνα με τον ορισμό της Ευρωπαϊκής Επιτροπής *«ανοιχτά δεδομένα είναι όσα διατίθενται ελεύθερα προς επαναχρησιμοποίηση, η οποία περιλαμβάνει χρήση των δεδομένων για σκοπούς που προβλέπονται ή όχι από τον πρωτότυπο δημιουργό»* [*Ευρωπαϊκή Επιτροπή*, 2011]. Τα σημαντικότερα χαρακτηριστικά των ανοιχτών δεδομένων, όπως καθορίζονται και από το Open Knowledge Foundation [https://okfn.org/opendata] είναι:

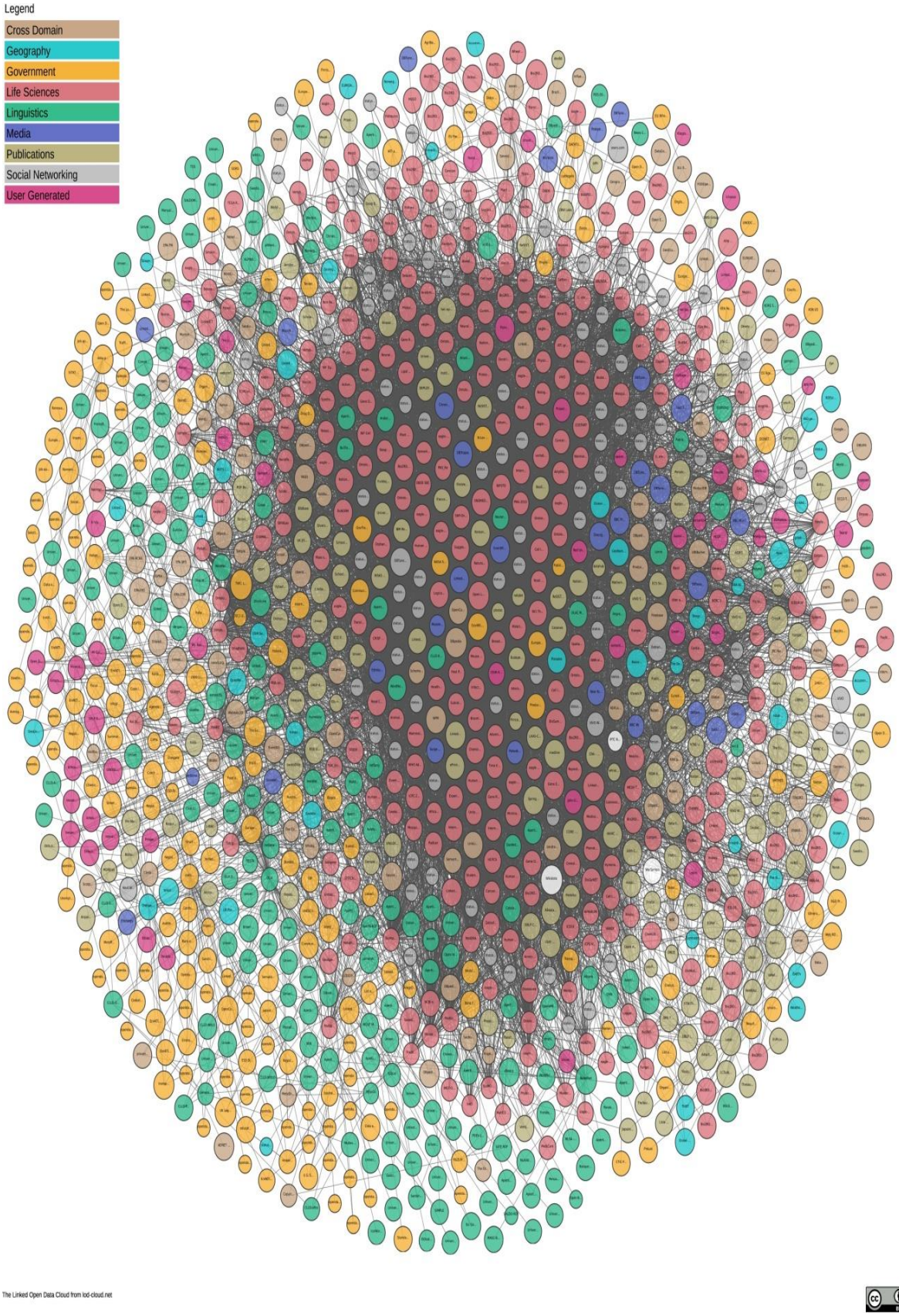
• *Η Διαθεσιμότητα και Προσβασιμότητα*: Τα δεδομένα πρέπει να διατίθενται αυτούσια, με χαμηλό κόστος αναπαραγωγής, να είναι διαθέσιμα στο Διαδίκτυο και προσπελάσιμα στους χρήστες και προσφέρονται σε κάποια μορφή πρακτικά αναγνώσιμη.
• *Η Επαναχρησιμοποίηση και Αναδιανομή*: Τα δεδομένα θα πρέπει να είναι διαθέσιμα υπό όρους που επιτρέπουν την επαναχρησιμοποίηση, την αναδιανομή και την ενδεχόμενη διασύνδεσή τους με άλλα σύνολα δεδομένων.
• *Η Καθολική Συμμετοχή*: Τα δεδομένα θα πρέπει να είναι διαθέσιμα προς χρησιμοποίηση, επαναχρησιμοποίηση και να αναδιανομή από οποιοδήποτε χρήστη και να υφίστανται διακρίσεις ανάλογα τον τομέα δραστηριότητας, τα πρόσωπα, τις ομάδες ή οποιαδήποτε άλλη διάκριση στη χρήση των δεδομένων.

Τα Ανοιχτά Συνδεδεμένα Δεδομένα (Linked Open Data) που εξετάζονται στο πλαίσιο της παρούσας διπλωματικής εργασίας συνιστούν δεδομένα ελεύθερα προσβάσιμα σε όλους, χωρίς τεχνικούς ή νομικούς περιορισμούς, με άδειες χρήσης και όρους χρήσης

που υπόκεινται μόνο στους νόμους περί επαναχρησιμοποίησης πληροφοριών του δημόσιου τομέα. Τα Ανοιχτά Συνδεδεμένα Δεδομένα δεν έχουν περιορισμούς, εφόσον τα δεδομένα παραμένουν άθικτα και δεν συνιστούν αντικείμενο εκμετάλλευσης, αναφέρεται η πηγή τους και καταγράφεται η ημερομηνία της τελευταίας επικαιροποίησής τους [*Stroetmann*, 2013].

Έχοντας κατανοήσει τη δυναμική και τα πλεονεκτήματα της διασύνδεσης της πληροφορίας, διάφοροι οργανισμοί, φορείς, κυβερνήσεις αλλά και μεμονωμένα άτομα άρχισαν να δημοσιεύουν τα δεδομένα τους στο Διαδίκτυο, με αποτέλεσμα τη δημιουργία του Ιστού Δεδομένων (Web of Data) [Linked Data, (n.d.). https://el.wikipedia.org/wiki/Linked_Data], ο οποίος με την σειρά του οδήγησε στη δημιουργία ενός γιγάντιου παγκόσμιου γράφου που αποτελείται από δισεκατομμύρια δηλώσεις. Η εικόνα 1 που ακολουθεί είναι το λεγόμενο διάγραμμα *LOD cloud* τον Ιούνιο του 2018, όπου απεικονίζονται όλα τα σύνολα δεδομένων (datasets), ή αλλιώς οι επονομαζόμενες φυσαλίδες. Τα χρώματα που εμφανίζονται είναι τα διαφορετικά πεδία στον Ιστό Δεδομένων. Οι συνδέσεις μεταξύ των φυσαλίδων παρουσιάζονται στο διάγραμμα ως βέλη. Η κατεύθυνση του βέλους υποδεικνύει το dataset που περιέχει του συνδέσμους ενώ τα αμφίδρομα βέλη υποδηλώνουν συνήθως ότι οι σύνδεσμοι αντικατοπτρίζονται και στα δύο datasets. Έως τον Ιούνιο του 2018 υπήρχαν 1224 σύνολα δεδομένων με 16.113 συνδέσεις.

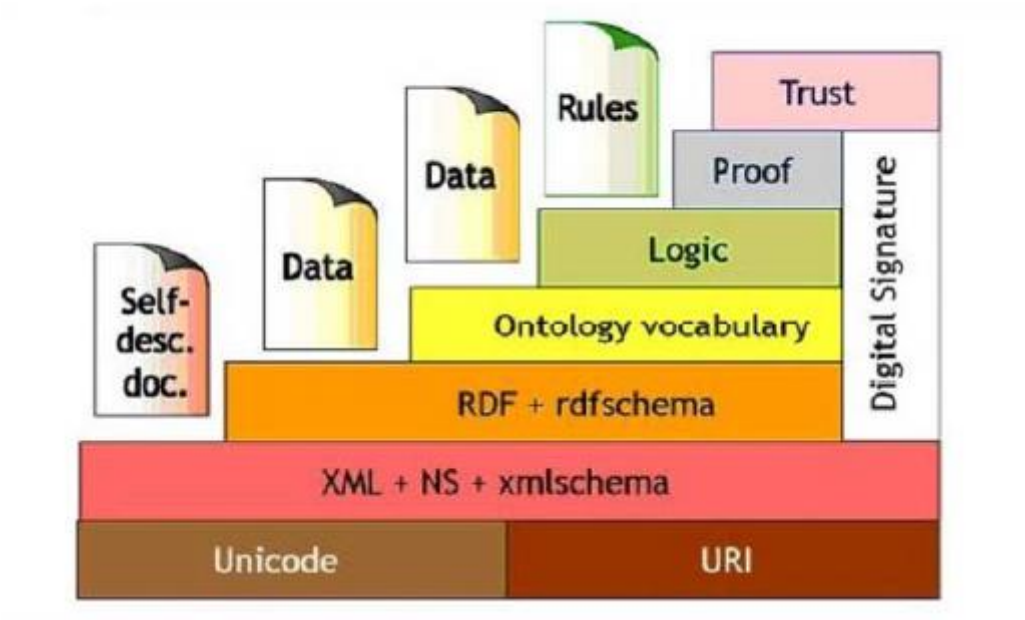
Εικόνα 1: LOD Cloud



2.2.2 Η εξέλιξη του Διαδικτύου μέχρι τη δημιουργία του Σημασιολογικού Ιστού

Οι πρωτογενείς δομικές μονάδες του διαδικτύου είναι το πρωτόκολλο HTTP (Hyper Text Transfer Protocol), ένα πρότυπο επικοινωνίας για τη μεταφορά εγγράφων, και τα έγγραφα σε γλώσσα HTML (Hyper Text MarkUp Language), τα οποία συνδέονται με υπερσυνδέσμους (hyperlinks) και αναγνωρίζονται μέσω του URL (Uniform Resource Locator) που καθορίζει την τοποθεσία τους στον Ιστό. Το γεγονός όμως ότι τα HTML έγγραφα δεν παρείχαν ικανοποιητικούς μηχανισμούς για την περιγραφή και σύνδεση μεταξύ τους, οδήγησαν στην επέκταση του Παγκόσμιου Ιστού με τον Σημασιολογικό Ιστό, όπου τα δεδομένα δημοσιεύονται δομημένα με την χρήση του μοντέλου RDF (Resource Description Framework) [*Klyne & Caroll*, 2004] και URI (Unified Resource Identifiers), έννοια ευρύτερη από αυτή του URL. Με τον τρόπο αυτό η πληροφορία αποκτά νόημα σαφώς προσδιορισμένο και γίνεται δυνατή η αμφίδρομη επικοινωνία μεταξύ ανθρώπου και υπολογιστή μέσω της κοινής γλώσσας που χρησιμοποιείται.

Εικόνα 2: Μορφή Σημασιολογικού Ιστού



Σύμφωνα με την εικόνα 2, στο κατώτερο επίπεδο του Σημασιολογικού Ιστού θα βρίσκεται το Unicode, το οποίο θα εξασφαλίζει την επικοινωνία ανάμεσα σε διαφορετικές γλώσσες και τα URI (Uniform Resource Identifier), στα οποία θα αναφερθούμε στην συνέχεια. Στο αμέσως επόμενο επίπεδο θα βρίσκεται η XML, η NS

και το XMLSchema ως κύρια γλώσσα έκφρασης στον ιστό. Αμέσως πιο πάνω η RDF (Resource Description Framework) και το RDFSchema ως κύρια γλώσσα μεταδεδομένων. Ακολουθεί το λεξιλόγιο οντολογιών και εν συνεχεία το λογικό επίπεδο ακολουθούμενο από το επίπεδο απόδειξης. Τέλος θα υπάρχει το επίπεδο της αξιοπιστίας.

Προκειμένου τα δεδομένα να γίνουν μέρος ενός ενιαίου παγκόσμιου χώρου δεδομένων, ο Tim Berners-Lee διατύπωσε το 2006 τους ακόλουθους κανόνες δημοσίευσής τους στο διαδίκτυο:

1.Χρήση URIs ως ονόματα πραγμάτων.
2. Χρήση HTTP URIs ώστε να μπορούν να αναζητηθούν από τον άνθρωπο.
3. Όταν κάποιος αναζητά ένα URI, πρέπει να του δίνεται χρήσιμη πληροφορία χρησιμοποιώντας καθιερωμένα πρότυπα όπως RDF, SPARQL.
4. Να περιλαμβάνονται σύνδεσμοι σε άλλα URIs, ώστε να μπορούν να ανακαλυφθούν περισσότερα πράγματα.

Ακολούθως, τον Απρίλιο του 2010 ο T.Berners-Lee διατύπωσε το σχήμα των πέντε αστέρων για τα ανοικτά δεδομένα μιας και γινόταν προσπάθεια από τις κυβερνήσεις να γίνει γνωστή η διάθεση των συνδεδεμένων δεδομένων για ελεύθερη χρήση (Linked Open Data). Το σχήμα δεδομένων παρουσιάζεται παρακάτω:

★ Διαθέσιμα στον Ιστό, σε οποιαδήποτε μορφή, αλλά με ανοικτή άδεια ώστε να είναι ανοιχτά δεδομένα.

★★ Διαθέσιμα ως δομημένα δεδομένα αναγνώσιμα από μηχανές (πχ excel αντί σαρωμένη εικόνα πίνακα).

★★★ Δομημένα δεδομένα αναγνώσιμα από μηχανές αλλά όχι σε ιδιόκτητες μορφές (π.χ. CSV αντί για excel).

★★★★ Όλα τα παραπάνω και επιπλέον ανοιχτών προτύπων του W3C (RDF, SPARQL) για να ταυτοποιούνται αντικείμενα.

★★★★★ Σύνδεση των δεδομένων με άλλα δεδομένα για την παροχή περιεχομένου.

2.2.3 Οι τεχνικές μορφοποίησης και δημοσίευσης των δεδομένων για την εξασφάλιση της διασύνδεσής τους στον Ιστό

Κατά την μελέτη της βιβλιογραφίας, προέκυψε περαιτέρω ότι για την εξασφάλιση της διαλειτουργικότητας (interoperability) μεταξύ των δεδομένων, έπρεπε να εφαρμοσθούν πρότυπα και τεχνικές στον τρόπο σχηματισμού τους ώστε να επιτραπεί η άντληση των απαιτούμενων κάθε φορά δεδομένων και τη επεξεργασία τους από οποιαδήποτε μηχανή. Τα εφαρμοζόμενα πρότυπα πρέπει να είναι ανεξάρτητα από τον τύπο ή το είδος της μηχανής ή της εφαρμογής που χρησιμοποιείται και τόσο ευέλικτα, ώστε οι νέες πληροφορίες για το ίδιο πεδίο γνώσης ή για τα νέα πεδία, να γίνονται άμεσα διαθέσιμες. Οι βασικές πρακτικές που ακολουθούνται σε τεχνικό επίπεδο για τη μορφοποίηση των δεδομένων σύμφωνα με την έρευνα των αντίστοιχων πηγών αναλύονται συνοπτικά στα εξής:

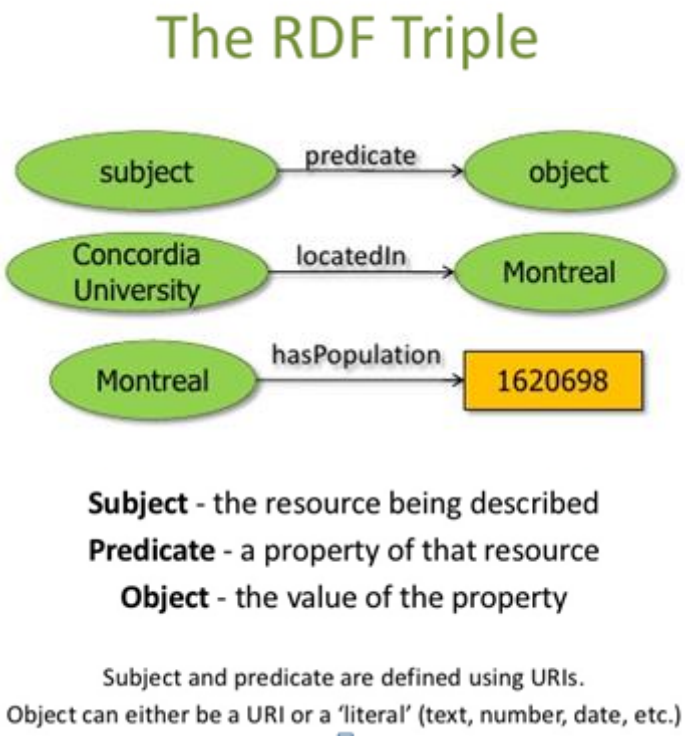
• ***Χρήση γλώσσας XML*:** Η XML (Extensible Markup Language) αποτελεί σημαντική τεχνολογία για την ενσωμάτωση του Σημασιολογικού Ιστού στο ήδη υπάρχον διαδίκτυο. Έχει γίνει αποδεκτή ως ένα ανεξάρτητο εργαλείο σύνταξης για τη μεταφορά δεδομένων μεταξύ προγραμμάτων που στο παρελθόν απευθυνόταν αποκλειστικά στη γλώσσα HTML και για να καλύψει τις αδυναμίες της τελευταίας. Ένα έγγραφο XML είναι δομημένο ιεραρχικά σε μορφή δέντρου, όπου τα στοιχεία περικλείονται μέσα σε άλλα στοιχεία και υπάρχει ένα βασικό στοιχείο ανώτερου επιπέδου, το οποίο περιέχει όλα τα υπόλοιπα. Κάθε στοιχείο αποτελείται από μια ετικέτα αρχής, το περιεχόμενο και μια ετικέτα τέλους. Στην XML ο συντάκτης πρέπει να καθορίσει τόσο τις ετικέτες όσο και τη δομή του εγγράφου. Το XML Schema είναι η γλώσσα που καθορίζει τη δομή των XML εγγράφων, αποτελεί δηλαδή το «συντακτικό» του XML κειμένου. Με τη χρήση της απλοποιείται η ανάγνωση, η μεταφορά και η διαθεσιμότητα των δεδομένων σε ανθρώπους, μηχανές, εφαρμογές κ.λπ.

• ***Το μοντέλο RDF*:** Το μοντέλο RDF (Resource Description Framework) είναι ένα μοντέλο δεδομένων το οποίο αναπαριστά την πληροφορία που σχετίζεται με τους πόρους στον Παγκόσμιο Ιστό, δηλαδή τα μεταδεδομένα που αφορούν συμπληρωματικές πληροφορίες όπως τίτλος, συντάκτης, ημερομηνία εγγράφου κ.α. Αποτελεί ένα από τα βασικά μοντέλα για την περιγραφή και αναπαράσταση των πληροφοριών μέσα στον Ιστό δεδομένων [Resource Description Framework (RDF). Concepts and Abstract Syntax. (n.d.). Ανάκτηση από https://www.w3.org/TR/2004/REC-rdf-concepts-20040210]. Τα κύρια στοιχεία του RDF μοντέλου είναι:

• *Resources [πόροι]:* Μπορούμε να θεωρήσουμε ως resource ένα αντικείμενο, ένα πράγμα στο οποίο αναφερόμαστε (π.χ. βιβλίο, ταινία, τραγούδι κλπ). Η αναφορά σε ένα resource γίνεται με τη χρήση ενός URI (Uniform Resource Identifier) που είναι είτε URL είτε οτιδήποτε άλλο μπορεί να προσδιορίσει μοναδικά ένα resource.
• *Properties [ιδιότητες]:* Ορίζουν τις ιδιότητες και τις σχέσεις με τις οποίες περιγράφονται τα resources (π.χ. τίτλος, διάρκειας, καλλιτέχνης κλπ). Και τα properties αναγνωρίζονται με τη χρήση URIs.
• *Statements [δηλώσεις]:* Δίνουν την τιμή μιας ιδιότητας για μια συγκεκριμένη πηγή. Υπάρχουν τρεις τρόποι αναπαράστασης των RDF statements: α) με τη χρήση τριπλέτων (triples), β) με τη χρήση κατευθυνόμενων γράφων με ετικέτες στις ακμές (directed labeled graphs) και γ) με τη χρήση ενός συντακτικού παρόμοιου με την XML.

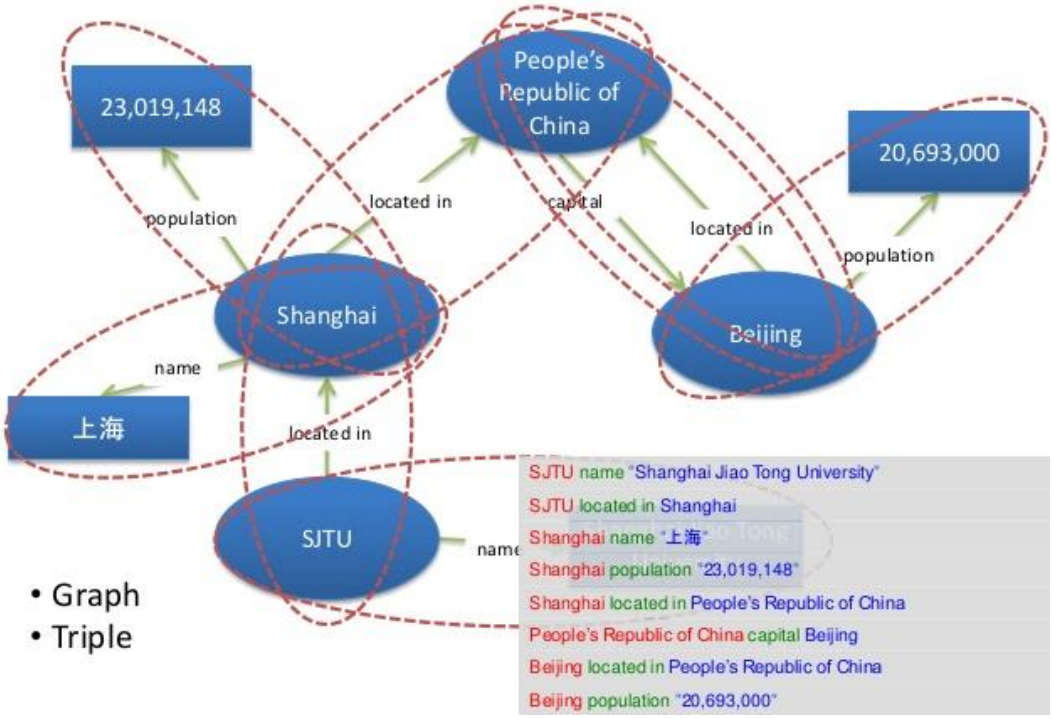
Στην αναπαράσταση με τη χρήση τριπλέτων [εικόνα 3], η δήλωση RDF αποτελείται από τρία επιμέρους συστατικά: τον πόρο/υποκείμενο (subject), την ιδιότητα/κατηγόρημα (predicate) και την τιμή ιδιότητας/αντικείμενο (object). Το υποκείμενο και το κατηγόρημα καθορίζονται με URIs, ενώ το αντικείμενο μπορεί να είναι είτε URI είτε literal (απόλυτη τιμή πχ. Κείμενο, αριθμός, ημερομηνία κλπ).

Εικόνα 3: Αναπαράσταση δήλωσης RDF με τριπλέτα



Στην περίπτωση που η δήλωση RDF αναπαρίσταται με γράφο [εικόνα 4], τότε έχουμε ένα κόμβο για το subject, έναν για το object και μια ακμή για το predicate με κατεύθυνση από το subject στο object. Οι RDF γράφοι μπορούν ακόμη να κωδικοποιηθούν σε διάφορες μορφές αναγνώσιμες από μηχανές, όπως είναι οι Turtle, RDFa, RDF/XML μιας και προσφέρουν μεγάλη ευελιξία στην προγραμματιστική επεξεργασία τους.

Εικόνα 4: Αναπαράσταση δήλωσης RDF με γράφο



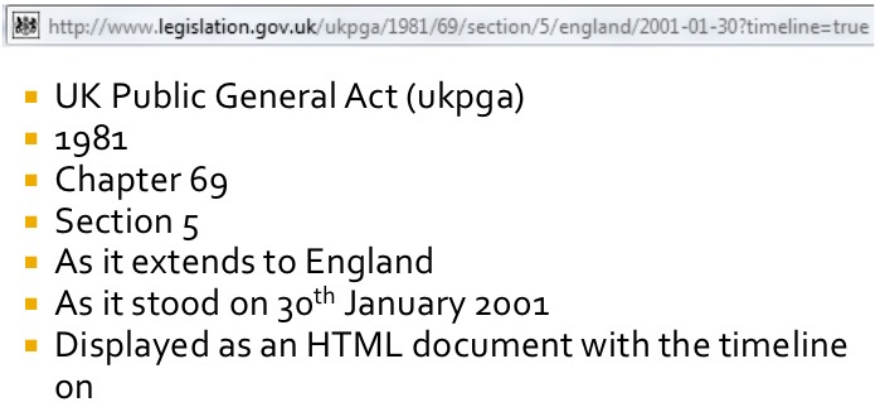
Η αναπαράσταση της RDF δήλωσης με τη χρήση ενός συντακτικού παρόμοιου με την XML[εικόνα 5] χρησιμοποιείται ώστε να είναι δυνατή η επεξεργασία των RDF δηλώσεων από τους υπολογιστές.

Εικόνα 5: Αναπαράσταση δήλωσης RDF της Οξφόρδης με XML

```
<?xml version="1.0" encoding="UTF-8"?>

<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:region="http://www.country-regions.fake/">

    <rdf:Description rdf:about="http://en.wikipedia.org/wiki/Oxford">
        <dc:title>Oxford</dc:title>
        <dc:coverage>Oxfordshire</dc:coverage>
        <dc:publisher>Wikipedia</dc:publisher>
        <region:population>10000</region:population>
        <region:principaltown rdf:resource="http://www.country-
        regions.fake/oxford"/>
    </rdf:Description>

</rdf:RDF>
```

• ***Τα URIs (Uniform Resource Identifiers)*:** Τα URI είναι μια συμπαγής ακολουθία χαρακτήρων που προσδιορίζει έναν αφηρημένο ή φυσικό πόρο (resource). Τα URI χρησιμεύουν ώστε τα αντικείμενα των RDF δηλώσεων να αναγνωρίζονται με μοναδικό τρόπο, δεδομένου ότι, εκτός από τον προσδιορισμό ενός πόρου, παρέχει και ένα μέσο εντοπισμού του, περιγράφοντας τον κύριο μηχανισμό πρόσβασης του (π.χ. τη θέση του δικτύου του). Το URI συγκεκριμένα χαρακτηρίζει αντικείμενα τα οποία είναι προσβάσιμα μέσω διαδικτύου όπως ηλεκτρονικά έγγραφα, εικόνες, υπηρεσίες, ή σύνολα από διάφορους άλλους πόρους. Επιπλέον χαρακτηρίζει αντικείμενα τα οποία δεν είναι προσβάσιμα μέσω του διαδικτύου, δηλαδή ανθρώπους ή οργανισμούς. Και τέλος, χαρακτηρίζει αφηρημένες έννοιες. Η επιλογή και η χρήση των κατάλληλων URIs είναι σημαντική για την παραγωγή Συνδεδεμένων Δεδομένων. Η επαναχρησιμοποίηση δημοφιλών URIs για την αναφορά σε οντότητες επιτρέπει, μεταξύ άλλων, τη σύνδεση ανάμεσα σε απομονωμένους γράφους RDF και διευκολύνει τη σημασιολογική διαλειτουργικότητα (semantic interoperability) των γράφων που προέρχονται από διαφορετικές πηγές. Σύμφωνα λοιπόν με το γενικό πλαίσιο των URI, το RDF χρησιμοποιεί αναφορές URI (uri references-URIs) για να αναγνωρίζει με μοναδικό τρόπο το υποκείμενο, το κατηγόρημα και το αντικείμενο μέσα σε μία τριπλέτα. Μία αναφορά URI αποτελείται από ένα URI και ένα αναγνωριστικό το οποίο μπαίνει στο τέλος του URI. Τέτοιο παράδειγμα είναι το URI http://www.example.org/countries#Italy το οποίο αποτελείται από το http://www.example.org/countries και το αναγνωριστικό Italy τα οποία διαχωρίζονται με τη δίεση (#). Οι αναφορές URI μπορούν να περιέχουν UNICODE ώστε τα URIrefs να χρησιμοποιούνται σε πολλές γλώσσες. Ως επί το πλείστον τα URIrefs βοηθούν το RDF ώστε να περιγράφει οποιοδήποτε δεδομένο και να εκφράζει τη μεταξύ τους σχέση. [https://www.w3.org/Provider/Style/URI].

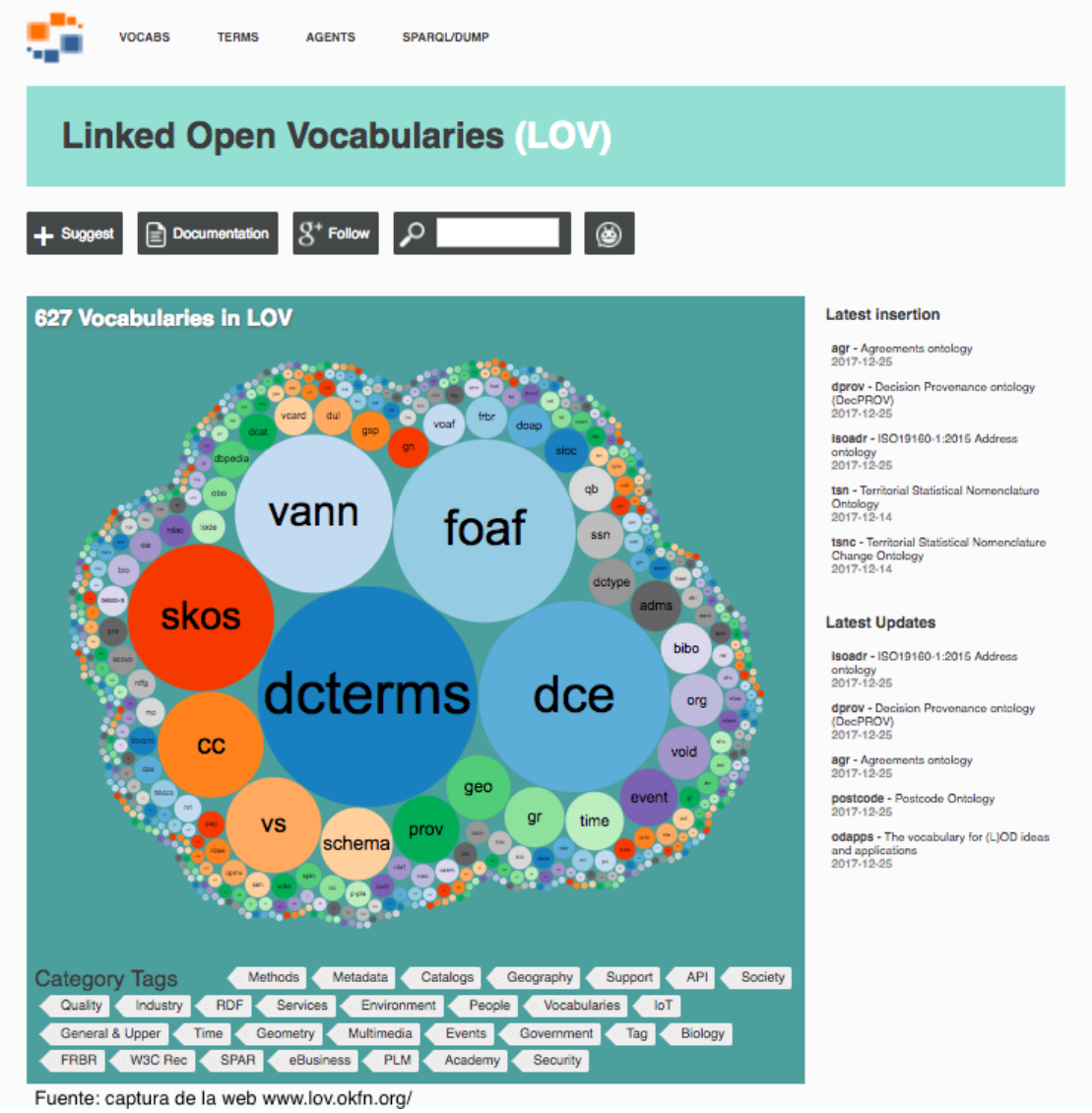
Παράδειγμα που απεικονίζει το URI του νόμου 1981/69 του Ηνωμένου Βασιλείου:
http://www.legislation.gov.uk/{***type***}/{***year***}/{***number***}[/{***section***}][/{***authority***}][/{***extent***}][/{***version***}]



• ***Ο ορισμός λεξιλογίου***: Κατά την πορεία της μελέτης έγινε αντιληπτό ότι για να καταστεί δυνατή η διασύνδεση μεταξύ των δεδομένων διαφόρων πηγών, όσοι επιθυμούν να δημοσιεύσουν δεδομένα πρέπει περαιτέρω να χρησιμοποιούν ένα ευρέως διαδεδομένο λεξιλόγιο. Δεδομένου ότι το Διαδίκτυο είναι ένα ανοιχτό περιβάλλον, διάφοροι πάροχοι πληροφοριών δημοσιεύουν δεδομένα που αφορούν την ίδια οντότητα του πραγματικού κόσμου, εισάγοντας όμως διαφορετικά URIs για τον προσδιορισμό της. Για παράδειγμα, το DBpedia χρησιμοποιεί το URI http://dbpedia.org/resource/Berlin για τον εντοπισμό του Βερολίνου, ενώ ο Geonames χρησιμοποιεί το URI http://sws.geonames.org/2950159/. Και τα δύο URI αναφέρονται στην ίδια παγκόσμια οντότητα (το Βερολίνο) και γι' αυτό αποκαλούνται ψευδώνυμα URI. Τα ψευδώνυμα URI είναι κοινά στον Παγκόσμιο Ιστό δεδομένων λόγω της ανομοιομορφίας του χρησιμοποιούμενου από κάθε πάροχο πληροφορίας URI που αναγνωρίζει την εκάστοτε οντότητα. Για να καταστεί λοιπόν δυνατή η διασύνδεση των πληροφοριών και άρα οι χρήστες αλλά και οι μηχανές να μπορούν να αναγνωρίσουν ότι και τα δύο παραπάνω URIs αφορούν στην ίδια οντότητα, τίθεται η σύνδεση μέσω της Γλώσσας Οντολογίας του Διαδικτύου (OWL) και πιο συγκεκριμένα τίθεται το *owl:sameAs* μεταξύ των ψευδώνυμων URI. Δεν αρκεί λοιπόν οι πάροχοι να δημοσιεύουν τα δεδομένα τους ως ανοιχτά, χρησιμοποιώντας ο καθένας το δικό του λεξιλόγιο. Συνεπώς, η τεχνική που επιτρέπει τη διασύνδεση μεταξύ των δεδομένων είναι η επαναχρησιμοποίηση όρων από ευρέως γνωστά και διαδεδομένα λεξιλόγια RDF όπως FOAF (Friend of a Friend), το λεξιλόγιο που χρησιμοποιούν οι βιβλιοθήκες FRBR (Functional Requirements for Bibliographic) Records SIOC (Semantically-Interlinked Online Communities Ontology), SKOS (Simple Knowledge Organization System), DOAP (Description of a project), Dublin Core κ.α όπου είναι δυνατόν, προκειμένου να διευκολυνθεί η επεξεργασία των

συνδεδεμένων δεδομένων μέσω των μηχανών και των εφαρμογών. Τα βασικότερα λεξιλόγια μπορούν ανακτηθούν από τα https://www.w3.org/standards/techs/rdfvocabs#w3c_all και https://lov.linkeddata.es/dataset/lov [εικόνα 6].

Εικόνα 6: Λεξιλόγια των Linked Open Data



Fuente: captura de la web www.lov.okfn.org/

•***Η Γλώσσα Οντολογίας Διαδικτύου*** (Web Ontology Language-OWL) είναι μια γλώσσα Σημασιολογικού Ιστού σχεδιασμένη να αντιπροσωπεύει πλούσιες και πολύπλοκες γνώσεις για πράγματα, ομάδες πραγμάτων και σχέσεις μεταξύ των πραγμάτων. Σύμφωνα με τον W3C.org αποτελεί μια γλώσσα που βασίζεται στη λογική, έτσι ώστε να μπορεί να αξιοποιηθεί από προγράμματα ηλεκτρονικών υπολογιστών. Τα έγγραφα OWL, γνωστά ως οντολογίες, μπορούν να δημοσιεύονται στον Παγκόσμιο Ιστό και μπορούν να αναφέρουν ή να αναφέρονται από άλλες οντολογίες OWL. Η OWL είναι μέρος της

τεχνολογίας του Σημασιολογικού Ιστού, το οποίο περιλαμβάνει RDF, RDFS, SPARQL κ.λπ. Με τη χρήση της OWL, τα διαφορετικά λεξιλόγια συνδέονται σημασιολογικά μεταξύ τους και τα διαφορετικά RDF γραφήματα ενώνονται σε ένα ενιαίο «σύννεφο δεδομένων» (data cloud), το οποίο επιτρέπει την ανταλλαγή και την αναζήτηση πληροφοριών από τους χρήστες. Με τον τρόπο αυτό έχει επιτευχθεί η διαλειτουργικότητα ανάμεσα σε πολυάριθμα και αυτόνομα λεξιλόγια ή συστήματα περιγραφής λεξιλογίων.

Λαμβάνοντας υπόψη τα παραπάνω, οι εκδότες των δεδομένων θα πρέπει να ορίζουν νέα ορολογία με βάση την προέλευση των δεδομένων μόνο εφόσον τα υπάρχοντα λεξιλόγια δεν περιέχουν τους απαιτούμενες όρους [*Bizer, Cyganiak & Heath*, 2007]. Αν οριστεί νέα ορολογία, τότε τα αντίστοιχα νέα URI θα πρέπει να ανακατευθύνονται μέσα στον Ιστό [*Berrueta & Phipps, 2008*], έτσι ώστε οι χρήστες να έχουν τη δυνατότητα να ανακτούν ορισμούς των οντολογιών RDF Schema ή OWL καθώς και να βρίσκουν αντιστοιχίσεις σε άλλα λεξιλόγια. Μεταξύ άλλων, θα πρέπει να ορίζονται σύνδεσμοι λεξιλογίου όπως είναι οι owl:equivalentClass και owl:equivalentProperty, οι οποίοι χρησιμοποιούνται για την ισοδυναμία δύο όρων, οι rdfs:subClassOf, rdfs:subPropertyOf χρησιμοποιούνται για να δηλώσουν τις ιεραρχίες, τέλος οι skos:broadMatch, and skos:narrowMatch χρησιμοποιούνται για να δηλώσουν μικρότερη ισοδυναμία μεταξύ των δεδομένων.

• ***Η σύνδεση μεταξύ των δεδομένων***: οι παραπάνω τεχνικές εφαρμόζονται προκειμένου να καταστεί στην συνέχεια δυνατή η σύνδεση των δεδομένων στον Ιστό, όπως ορίζει η τελευταία αρχή και το 5^{ο} αστέρι του Tim Berners-Lee. Η σύνδεση των δεδομένων στον Ιστό διακρίνεται σε εσωτερική και εξωτερική. Η εσωτερική σύνδεση υπάρχει όταν συνδέονται οι πόροι μεταξύ τους μέσα στην ίδια πηγή και η εξωτερική υπάρχει όταν οι πόροι που συνδέονται μεταξύ τους είναι σε διαφορετικές πηγές. Όπως προεκτέθηκε, η σύνδεση των δεδομένων γίνεται μέσα σε μία τριπλέτα RDF και οι πόροι είναι το υποκείμενο και το αντικείμενο οι οποίοι παρουσιάζονται με τα αντίστοιχα URIs τους, ενώ το κατηγόρημα συνδέει τους δύο προηγούμενους πόρους επίσης με ένα URI. Μπορούμε να διακρίνουμε τρία είδη συνδέσεων: α) σύνδεσμοι σχέσης που συνδέουν δύο οντότητες σε πραγματικό χρόνο, όπως για παράδειγμα ο χρόνος γέννησης ενός ατόμου, β) σύνδεσμοι ταυτότητας που συνδέουν δύο URIs τα οποία υπάρχουν σε διαφορετικά σύνολα δεδομένων ώστε να δείξουν ότι αναφέρονται στον ίδιο πόρο (πχ το γνώρισμα «http://www.w3.org/2002/07/owl#sameAs» χρησιμοποιείται γι' αυτό το λόγο από την Web Ontology Language) και γ) σύνδεσμοι λεξιλογίου που χρησιμοποιούνται για την

εννοιολογική σύνδεση των δεδομένων, έτσι ώστε να δημιουργείται νέα γνώση από τα συμπεράσματα που προκύπτουν κάθε φορά [*van Assem, M, 2010*]. Η δημιουργία συνδέσμων RDF ανάμεσα στους πόρους από δύο σύνολα δεδομένων γίνεται με δύο τρόπους: α) χειροκίνητα ή β) αυτόματα. Η αυτόματη σύνδεση των πόρων μπορεί να γίνει είτε με τη χρήση ενός SPARQL endpoint είτε με φυλλομετρητή Συνδεδεμένων Δεδομένων, αλλά και με URIs ευρετήρια όπως είναι το Sindice και το Falcons, με τη βοήθεια των οποίων γίνεται η σύνδεση των URIs με τη διαδικασία της αναζήτησης συγκεκριμένων λέξεων κλειδιών. Από την άλλη πλευρά η χειροκίνητη σύνδεση εφαρμόζεται για πολύ μικρή ποσότητα δεδομένων και όχι για εκατομμύρια URIs από διαφορετικές πηγές. Επίσης, με τη χειροκίνητη σύνδεση συνήθως συνδέεται ένα URI με το κείμενο το οποίο περιγράφει το αντικείμενο και όχι με το URI αυτό κάθε αυτό.

• ***Η πρόσβαση στα Συνδεδεμένα Δεδομένα***: Εφόσον εφαρμοστούν τα ανωτέρω πρότυπα και τεχνικές, ο χρήστης και οι μηχανές αποκτούν πρόσβαση στα Συνδεδεμένα Δεδομένα με τους εξής τρόπους: α) μέσω ενός Linked Data Browser που αποτελεί μία εφαρμογή παρόμοια με τους Web Browsers. Ο χρήστης μπορεί να πλοηγηθεί σε RDF συνδέσμους μέσω ιστοσελίδων. Ο χρήστης πρέπει να χρησιμοποιήσει λέξεις κλειδιά για να προβεί σε αναζήτηση που αυτή με τη σειρά της γίνεται από μηχανές αναζήτησης οι οποίες περνούν από τον ιστό δεδομένων σε συνδέσμους RDF, επιπλέον ενοποιούν και φιλτράρουν τα δεδομένα που βρίσκουν με τη χρήση των τάξεων. Γνωστοί linked data browsers είναι οι Tabulator, Disco, Ontology-browser, Falcons Explorer, β) μέσω του SPARQL endpoint που εμφανίζει πολλές ομοιότητες με τη γλώσσα προγραμματισμού SQL, γ) μέσω λήψης των δεδομένων σε τοπικό αρχείο και δ) μέσω εφαρμογών ειδικού σκοπού που προσφέρουν στο χρήστη συγκεκριμένα δεδομένα ενός πεδίου με συγκεκριμένο τρόπο.

• ***Τα SPARQL endpoint:*** Τα SPARQL endpoint είναι σημεία στον Ιστό που παρέχουν πρόσβαση σε RDF δεδομένα μέσω του SPARQL Protocol and RDF Query Language (SPARQL). Η SPARQL είναι μια σημασιολογική γλώσσα ερωτήσεων για βάσεις δεδομένων που συνεργάζεται με προαναφερθείσες τεχνολογίες του Σημασιολογικού Ιστού (RDF, RDF Schema, OWL κλπ) και έχει τη δυνατότητα να ανακτά και να επεξεργάζεται δεδομένα που είναι αποθηκευμένα σε αποθετήρια τριάδων RDF. Το SPARQL ερώτημα που παράγεται λαμβάνει υπόψη του το γεγονός ότι η απάντηση μπορεί να συνδυάζει πληροφορίες που βρίσκονται κατανεμημένες σε διαφορετικά εναποθετήρια δεδομένων, ακόμη και μη αναμενόμενα, και αν υπάρχει διαθέσιμη διεπαφή (interface) για αυτά, τότε αρκεί να την τροφοδοτήσουμε με το ερώτημα για να λάβουμε την απάντηση.

Περαιτέρω, ο όρος «τελικό σημείο» (endpoint) έχει μια γενικότερη έννοια. Κατά τον ορισμό της Κοινοπραξίας του W3C [https://www.w3.org/TR/ws-addr-core/#eprinfomodel], το endpoint είναι μια προδιαγραφή του Σημασιολογικού Ιστού που καθορίζει ένα βασικό σύνολο ιδιοτήτων, παρέχοντας παράλληλα τη δυνατότητα και σε άλλες προδιαγραφές να επεκτείνουν ή / και να προσθέσουν ιδιότητες. Το endpoint αποτελείται από τις ακόλουθες αφηρημένες ιδιότητες: α) τη διεύθυνση του τελικού σημείου, β) τις επιμέρους παραμέτρους που συνδέονται με το τελικό σημείο προκειμένου να διευκολύνουν τη διαδραστικότητα και γ) τα μεταδεδομένα. Τα αποτελέσματα των ερωτημάτων που τίθενται μέσω του SRARQL endpoint επιστρέφονται σε μία ή περισσότερες μορφές επεξεργασμένες από μηχανή και άρα τόσο η διατύπωση των ερωτημάτων όσο και η ανθρώπινη αναγνώσιμη παρουσίαση των αποτελεσμάτων θα πρέπει τυπικά να υλοποιούνται από το καλούμενο λογισμικό και να μην γίνονται χειροκίνητα από ανθρώπους [https://en.wikipedia.org/wiki/SPARQL].

• ***Τα Μετα-δεδομένα (metadata)***: Προκειμένου να αυξηθεί ακόμη περισσότερο η χρησιμότητα των Συνδεδεμένων Δεδομένων για τους χρήστες, αυτά θα πρέπει να δημοσιεύονται μαζί και με διάφορους τύπους μετα-δεδομένων [*Hartig, 2009*]. Οι βασικές πληροφορίες μπορούν να παρασχεθούν με την χρήση των όρων του λεξιλογίου Dublin Core ή του λεξιλογίου Εκδόσεων του Σημασιολογικού Ιστού [*Carroll, Bizer, Hayes& Stickler,* 2005]. Τα μετα-δεδομένα περιλαμβάνουν πληροφορίες για τα δεδομένα στα οποία αναφέρονται και σε αρκετές περιπτώσεις είναι κωδικοποιημένα. Τα μετα-δεδομένα μπορεί να είναι διαχειριστικά (πότε κι από ποιόν δημιουργήθηκαν), περιγραφικά (τι περιλαμβάνει το σύνολο δεδομένων), τεχνικά (κάποιο πρότυπο που ακολουθούν τα δεδομένα, του τρόπους πρόσβασης σε αυτά), χρήσης (κάτω από ποια άδεια χρήσης διατίθενται, πνευματικά δικαιώματα), δεδομένα ευρετηρίου (λέξεις κλειδιά που θα διευκολύνουν την εύρεση των δεδομένων από τις μηχανές αναζήτησης) κ.ο.κ. Όσο πιο πλήρης είναι η περιγραφή των πόρων που προσφέρουν τα συνδεδεμένα δεδομένα, τόσο πιο πλήρη είναι και τα αντίστοιχα μετά-δεδομένα. Εκτός από πληρότητα, ακρίβεια, συνέπεια, τα μετα-δεδομένα θα πρέπει να μπορούν να αναγνωσθούν και να τύχουν μηχανικής επεξεργασίας.

• ***Τα εργαλεία δημοσίευσης:*** Μέχρι σήμερα έχει αναπτυχθεί μια μεγάλη γκάμα εργαλείων δημοσίευσης συνδεδεμένων δεδομένων. Τα εργαλεία αυτά είτε μετατρέπουν το περιεχόμενο των αποθηκών RDF σε Συνδεδεμένα δεδομένα στον Ιστό είτε παρέχουν συνδέσεις με δεδομένα που προέρχονται από παλαιότερες πηγές που δεν χρησιμοποίησαν

ακόμη το πρότυπο RDF. Τα εργαλεία που αναφέρονται αμέσως παρακάτω διασφαλίζουν ότι τα δεδομένα δημοσιεύονται ακολουθώντας τις βέλτιστες πρακτικές της κοινότητας για τα Συνδεδεμένα Δεδομένα [*Sauermann & Cyganiak*, 2008]. Όλα τα εργαλεία δημοσίευσης υποστηρίζουν την αποδιαμόρφωση των URIs σε δηλώσεις RDF. Επιπλέον, μερικά από τα εργαλεία παρέχουν πρόσβαση σε ερωτήματα SPARQL.

Παραδείγματα εργαλείων δημοσίευσης συνδεδεμένων δεδομένων είναι:

• **Ο διακομιστής D2R Server** είναι ένα εργαλείο μέσω του οποίου είναι δυνατή η δημοσίευση μη σχεσιακών RDF βάσεων δεδομένων ως Συνδεδεμένα Δεδομένα στον Ιστό. Αυτό επιτυγχάνεται χρησιμοποιώντας μια μέθοδο αντιστοίχισης μεταξύ του σχεσιακού σχήματος της βάσης δεδομένων και του λεξιλογίου RDF, επιτρέποντας την αναζήτηση δεδομένων μέσω του πρωτοκόλλου SPARQL.

• **Virtuoso Universal Server**. Ο διακομιστής OpenLink Virtuoso [Endnote: http: //www.openlinksw.com/dataspace/dav/wiki/Main/VOSRDF] παρέχει δεδομένα RDF μέσω μιας διεπαφής Συνδεδεμένων Δεδομένων και ενός SPARQL endpoint. Τα δεδομένα RDF μπορούν είτε να αποθηκευτούν απευθείας στο Virtuoso είτε να δημιουργηθούν με ασφάλεια από μη σχεσιακές βάσεις δεδομένων RDF με βάση μια χαρτογράφηση.

• **The Triplify toolkit** [*Auer et al, 2009*] που δίνει τη δυνατότητα στους προγραμματιστές να επεκτείνουν τις υφιστάμενες εφαρμογές Web με τα αρχικά σύνολα των Συνδεδεμένων Δεδομένων. Βασισμένο σε πρότυπα ερωτημάτων γλώσσας SQL, το εργαλείο Triplify εξυπηρετεί την προβολή Συνδεδεμένων δεδομένων στη βάση δεδομένων της εφαρμογής.

• **Το SparqPlug** [*Coetzee, Heath& Motta, 2008*] είναι μια υπηρεσία που επιτρέπει την εξαγωγή των Συνδεδεμένων Δεδομένων από παλαιότερα έγγραφα HTML στον Ιστό που δεν περιέχουν πληροφορίες RDF. Η υπηρεσία λειτουργεί με τη σειριακή τοποθέτηση του HTML DOM ως RDF, επιτρέποντας στους χρήστες να καθορίσουν τα ερωτήματα SPARQL που μετασχηματίζουν τα στοιχεία αυτού σε ένα γράφο RDF της επιλογής τους

• **Το έργο SIOC** (Semantically Interlinked Online Communities – Σημασιολογικά Διασυνδεδεμένες Κοινότητες του Διαδικτύου) που έχει αναπτύξει το περικάλυμμα των συνδεδεμένων δεδομένων για πολλαπλούς μηχανισμούς blogging, συστήματα διαχείρισης περιεχομένου και φόρουμ συζητήσεων όπως WordPress, Drupal και phpBB [Endnote: http: //sioc-project.org/exporters].

2.2.4 Ανακεφαλαίωση

Από την βιβλιογραφική επισκόπηση των διαθέσιμων πηγών για τα ανοιχτά συνδεδεμένα δεδομένα, συμπεραίνεται ότι για να μπορούν οι εφαρμογές συνδεδεμένων δεδομένων να ανακαλύψουν σύνολα δεδομένων και να διευκολύνουν την ενσωμάτωση πληροφοριών από πολλαπλές πηγές, οι εκδότες συνδεδεμένων δεδομένων θα πρέπει να συμμορφώνονται με μια σειρά βέλτιστων πρακτικών [*Heath & Bizer, 2011*], οι οποίες ομαδοποιούνται ως εξής:

Σύνδεση: Με τον ορισμό δεσμών RDF, οι πάροχοι δεδομένων συνδέουν τα σύνολα δεδομένων τους με ένα παγκόσμιο γράφημα δεδομένων, το οποίο μπορεί να πλοηγηθεί από εφαρμογές και επιτρέπει την εξακρίβωση πρόσθετων δεδομένων ακολουθώντας συνδέσμους RDF.

Χρήση λεξιλογίου: οι εκδότες δεδομένων πρέπει να χρησιμοποιούν όρους από ευρέως γνωστά και χρησιμοποιούμενα λεξιλόγια για να διευκολύνουν την ερμηνεία των δεδομένων τους. Εάν οι πάροχοι δεδομένων χρησιμοποιούν τα δικά τους λεξιλόγια, οι όροι αυτών των ιδιόκτητων λεξιλογίων πρέπει να μπορούν να μεταφερθούν στο πλαίσιο των σχημάτων RDF ή OWL. Οι ορισμοί των ιδιόκτητων όρων λεξιλογίου πρέπει να περιέχουν συνδέσμους RDF από ευρέως χρησιμοποιούμενα λεξιλόγια για να διευκολύνουν την ερμηνεία τους.

Παροχή Μετα-δεδομένων: Τα Συνδεδεμένα Δεδομένα θα πρέπει να είναι όσο το δυνατόν πιο περιγραφικά και να περιλαμβάνουν μετα-δεδομένα, δηλαδή πληροφορίες που συνήθως δεν είναι ορατές στον τελικό χρήστη. Μία σημαντική μορφή μετα-δεδομένων είναι η προέλευση, η οποία καθορίζει την πηγή των συνόλων δεδομένων και επιτρέπει στις εφαρμογές να αξιολογούν την ποιότητά τους. Εάν τα σύνολα δεδομένων είναι προσβάσιμα και μέσω πρόσθετων μεθόδων πρόσβασης, όπως ένα τελικό σημείο SPARQL, τότε τα αρχεία των μεταδεδομένων θα περιέχουν πληροφορίες σχετικά με αυτές τις μεθόδους πρόσβασης [*Schmachtenberg, Bizer & Paulheim, 2014*].

2.3 ΤΑ ΑΝΟΙΧΤΑ ΣΥΝΔΕΔΕΜΕΝΑ ΝΟΜΙΚΑ ΔΕΔΟΜΕΝΑ (LEGAL LINKED OPEN DATA)

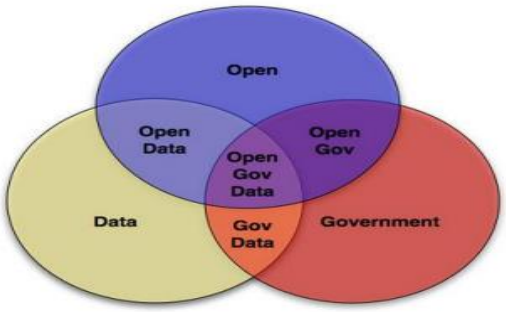
2.3.1 Από τα ανοιχτά κυβερνητικά δεδομένα στα ανοιχτά νομικά δεδομένα

Όπως προαναφέρθηκε, κατά τη διάρκεια εκπόνησης της παρούσας εργασίας και λαμβάνοντας ως αφορμή ένα άρθρο του Enrico Francesconi που δημοσιεύθηκε στις 5 Ιουνίου 2018, η βιβλιογραφική μας έρευνα επικεντρώθηκε στα ανοιχτά συνδεδεμένα

νομικά δεδομένα που αποτελούν ουσιαστικά επιμέρους κατηγορία κυβερνητικών δεδομένων καθότι παράγονται – νομοθετούνται – εκδίδονται από κοινοβούλια, αρμόδια κυβερνητικά όργανα ή δημόσιους λειτουργούς (δικαστές).

Η προσέγγιση των Συνδεδεμένων Δεδομένων δεν μεταβάλλει την τυπική αρχιτεκτονική στην οποία βασίζεται η φιλοσοφία του Σημασιολογικού Ιστού αλλά προωθεί τη δημοσίευση και διαχείριση δεδομένων χρησιμοποιώντας URI για την ταυτοποίησή τους και RDF για την περιγραφή τους χρησιμοποιώντας την υπάρχουσα XML οργάνωση των πόρων. Λαμβάνοντας υπόψη τις αρχές δημοσίευσης των δεδομένων στο Διαδίκτυο, το πρότυπο 5 αστέρων του Tim Berners-Lee και τα πλεονεκτήματα που παρέχει η ελεύθερη πρόσβαση στα κυβερνητικά δεδομένα αναπτύχθηκαν αρκετές πρωτοβουλίες ανοιχτής διακυβέρνησης σε ΗΠΑ (http://www.data.gov), Ηνωμένο Βασίλειο (http://www.data.gov.uk) αλλά και στην Ελλάδα (http://www.data.gov.gr, http://www.diavgeia.gov.gr). Τέτοιας φύσεως πρωτοβουλίες προωθούν το άνοιγμα των δημόσιων δεδομένων (που δημοσιεύονται υπό άδειες ελεύθερες από τη φύση τους) ως μέσο διαφάνειας και συμμετοχής των πολιτών στο δημοκρατικό σύστημα. Από αυτή την άποψη, οι δημόσιες διοικήσεις (ως εκδότες δεδομένων) είναι υπεύθυνες για την παροχή δεδομένων στο βέλτιστο δυνατό επίπεδο διαλειτουργικότητας, αφήνοντας στους πολίτες και άλλους δημόσιους και ιδιωτικούς φορείς (ως καταναλωτές δεδομένων) τη δυνατότητα εκμετάλλευσης και εμπλουτισμού των δεδομένων.

Εικόνα 7: Ανοιχτά Κυβερνητικά Δεδομένα



Ένα σημαντικό είδος κυβερνητικών δεδομένων είναι τα δεδομένα που σχετίζονται με τη νομοθεσία. Η νομοθεσία εφαρμόζεται σε κάθε πτυχή της ζωής των ανθρώπων και εξελίσσεται συνεχώς, δημιουργώντας ένα τεράστιο δίκτυο διασυνδεδεμένων νομικών εγγράφων. Ως εκ τούτου, είναι σημαντικό μια κυβέρνηση να προσφέρει υπηρεσίες που καθιστούν τη νομοθεσία εύκολα προσβάσιμη στο ευρύ κοινό, με στόχο την ενημέρωσή του, την υπεράσπιση των δικαιωμάτων του και την χρήση της νομοθεσίας ως μέρος της

δουλειάς του. Λαμβάνοντας υπόψη τα ανωτέρω το Ηνωμένο Βασίλειο, σε συνέχεια της πρωτοβουλίας για ανοιχτά κυβερνητικά δεδομένα που προαναφέρθηκε, δημιούργησε την ιστοσελίδα www.legislation.gov.uk, μια διαδικτυακή υπηρεσία που περιέχει όλη τη νομοθεσία που εκδόθηκε από το 1267 μέχρι και σήμερα, συμπεριλαμβανομένης της εξέλιξης και των τροποποιήσεων που έχει υποστεί με την πάροδο του χρόνου, καθώς και τους ισχύουσες νομοθετικές διατάξεις. Τα έγγραφα της ως άνω ιστοσελίδας μπορούν να προβληθούν σε διαφορετικές μορφές (PDF και XML), ενώ οι χρήστες έχουν επιπλέον τη δυνατότητα να επιλέξουν οποιονδήποτε τύπο νομοθεσίας και γεωγραφική περιοχή προκειμένου να εμφανιστούν ποιοι τύποι νόμων ισχύουν τοπικά. Από τεχνική άποψη, το www.legislation.gov.uk παρέχει προδιαγραφές API (Application Programming Interface) για πλήρη και ελεύθερη πρόσβαση στα υποκείμενα δεδομένα, ώστε να μπορούν να επαναχρησιμοποιηθούν υπό την άδεια ανοικτής κυβέρνησης (http: /data.gov.uk) [*Francesconi, 2018*].

Ωστόσο, διαπιστώθηκε ότι η δημοσίευση και η χρήση των νομικών πληροφοριών εμφανίζει εγγενείς δυσκολίες διότι: α) η εκάστοτε νομοθεσία και νομολογία παράγεται από διαφορετικά κόμματα, κοινοβούλια, υπουργεία, κυβερνητικά γραφεία, διάφορα δικαστήρια και ερευνητικούς οργανισμούς, προέρχεται δηλαδή από πολλές πηγές, β) το περιεχόμενο είναι ετερογενές και παράγεται χρησιμοποιώντας διαφορετικά εργαλεία, μορφές δεδομένων και πρακτικές, γ) οι δεσμοί μεταξύ των εγγράφων είναι συχνά ανεπίσημοι ή / και δεν καθίστανται σαφείς, δ) ο νόμος εν γένει είναι μια δυναμική, μεταβαλλόμενη οντότητα που υπόκειται διαρκώς σε προσθήκες και τροποποιήσεις. Συνεπώς, είναι σημαντικό να μπορεί ο χρήστης να ανατρέχει για παράδειγμα σε διαφορετικές εκδοχές ενός νόμου όπως ίσχυε μια συγκεκριμένη χρονική στιγμή, ε) συχνά η ερμηνεία της νομοθεσίας από τα δικαστήρια είναι πολύπλοκη και υπάρχουν παραπομπές σε πολλές διαφορετικές νομικές διατάξεις, στ) τα νομικά κείμενα έχουν κατά βάση εξειδικευμένη γλωσσική διατύπωση [*Boella et al., 2015*].

Αυτές οι προκλήσεις μπορούν να αντιμετωπιστούν μέσω της χρήσης της τεχνολογίας των συνδεδεμένων δεδομένων, ούτως ώστε τα νομικά δεδομένα να δημοσιεύονται όχι μόνο ως ακατέργαστα δεδομένα (raw data) αλλά υπό τη μορφή έτοιμων προς χρήση υπηρεσιών για τους τελικούς χρήστες και τις μηχανές [*Frosterus, Tuominen & Hyvonen, 2014*].

Τα βασικότερα πλεονεκτήματα της ελεύθερης πρόσβασης σε κυβερνητικά και νομικά δεδομένα συνοψίζονται στον πίνακα 1.

Πίνακας 1: Πλεονεκτήματα ανοιχτών κυβερνητικών και νομικών δεδομένων

ΠΛΕΟΝΕΚΤΗΜΑΤΑ	
Ανοιχτών Κυβερνητικών δεδομένων	Ανοιχτών Νομικών Δεδομένων *[Oksanen et al., 2018]*
• Διασφάλιση της διαφάνειας των κυβερνητικών ενεργειών ιδίως όσον αφορά στις κρατικές δαπάνες •Δημοκρατική υπευθυνότητα λόγω ευρύτατης ενημέρωσης των πολιτών για τις κυβερνητικές ενέργειες •Ενθάρρυνση δημοκρατικής συμμετοχής των πολιτών κατά τη νομοθετική διαδικασία (διαβούλευση). •Ενίσχυση των συνταγματικών δικαιωμάτων των πολιτών, όπως η συμμετοχή στην κοινωνία της πληροφορίας •Δημιουργία εμπιστοσύνης στην εκάστοτε κυβέρνηση • Προώθηση μεγαλύτερης λογοδοσίας • Εξάλειψη της διαφθοράς εκθέτοντάς την πιο εύκολα όταν λαμβάνει χώρα •Τήρηση της νομιμότητας και της χρηστής διοίκησης • Δημιουργία όλων των διοικητικών πράξεων σε μορφές με εύκολη πρόσβαση, πλοήγηση και κατανόηση, ανεξάρτητα από το επίπεδο γνώσης του πολίτη από τις εσωτερικές διαδικασίες της διοίκησης	• Οι Πύλες πληροφοριών. • Παροχή εξειδικευμένων πληροφοριών από τα μέσα μαζικής ενημέρωσης στο ευρύ κοινό. • Δημιουργία νομικών υπηρεσιών online που παρέχουν νομικές πληροφορίες κατά κύριο λόγο για επαγγελματίες του δικαίου, (π.χ. η Suomen Laki και η Edilex7 στη Φιλανδία). • Ορθότερη νομοθετική διατύπωση κατόπιν εξέτασης προγενέστερων διατάξεων προς αποφυγή αποκλίσεων. • Επεξεργασία και δημοσίευση των νομοθετικών συνόλων δεδομένων. • Ανάπτυξη εφαρμογών για παροχή ευφυιών υπηρεσιών που θα δύνανται να συλλέγουν και να ερμηνεύουν τα νομοθετικά έγγραφα ως σημασιολογικά δεδομένα, με σκοπό την κατανόηση του δικαίου από τους πολίτες. • Διευκόλυνση της έρευνας στον τομέα της νομοθεσίας και της νομικής πρακτικής ως μέθοδος ανάλυσης δεδομένων. • Παραγωγή περιεχομένου

2.3.2 Οι πρωτοβουλίες για την ελεύθερη πρόσβαση σε νομοθεσία και νομολογία και την υλοποίηση της διασύνδεσης των νομικών δεδομένων

Από την εκτενή έρευνα της διαθέσιμης βιβλιογραφίας προέκυψε ότι οι κυριότερες πρωτοβουλίες που αναλήφθηκαν ήταν οι εξής:

• *Το Eur-lex* [https://eur-lex.europa.eu], το οποίο περιέχει το σύνολο της ευρωπαϊκής νομοθεσίας με παραπομπές και σε προγενέστερες εκδόσεις διεθνών συνθηκών, οδηγιών και κανονισμών, τις προπαρασκευαστικές ενέργειες (αιτιολογικές εκθέσεις), τις διεθνείς συμφωνίες, αλλά και τις δικαστικές αποφάσεις που έχουν εκδοθεί από τα αρμόδια ευρωπαϊκά δικαστήρια, με παράλληλη δυνατότητα αναζήτησης σε πολλές βάσεις δεδομένων [http://eur-lex.europa.eu/n-lex/index_el].

• Οι ιστότοποι https://www.legifrance.gouv.fr και https://www.courdecassation.fr στη Γαλλία, http://www.gesetze-im-internet.de στη Γερμανία, https://www.supremecourt.uk στο Ηνωμένο Βασίλειο αλλά και τα http://www.et.gr, http://www.e-nomothesia.gr και https://www.areiospagos.gr στην Ελλάδα.

• *Το πρόγραμμα «NormeInRete» (NIR)* του Υπουργείου Δικαιοσύνης της Ιταλίας, για την υλοποίηση του οποίου υιοθετήθηκαν οι αρχές του Σημασιολογικού Ιστού και πιο συγκεκριμένα χρησιμοποιήθηκαν URIs για την αναγνώριση εγγράφων (URN-NIR), καθώς και πρότυπα XML (XML-NIR) και RDF / OWL για την αντιπροσώπευση της δομής και της σημασιολογίας των περιεχομένων των νομικών εγγράφων [*Francesconi*, 2006, *Spinosa, Francesconi, Lupo*, 2018].

• *Η πρωτοβουλία της Ομοσπονδιακής Γερουσίας της Βραζιλίας*, η οποία υιοθέτησε το μοντέλο URN και ανέπτυξε μια υπηρεσία επιβολής του νόμου (LexML Brazil) που εμπνεύστηκε από το NormeInRete καθότι βασίστηκε στο πρότυπο URN-LEX για την κατασκευή αναγνωριστικών εγγράφων.

• *Η υπηρεσία Zotero* (http://www.zotero.org) στην Ιαπωνία, η οποία αποτελεί μια ελεύθερη και ανοικτού κώδικα υπηρεσία ικανή να αναγνωρίζει αναφορές που προέρχονται είτε από βάσεις δεδομένων νομικών εκδοτών είτε απευθείας από τους χρήστες και να δημιουργεί βιβλιοθήκες με επιστημονικά άρθρα, επιτρέποντας παράλληλα τη διαλειτουργικότητα με άλλες υπηρεσίες.

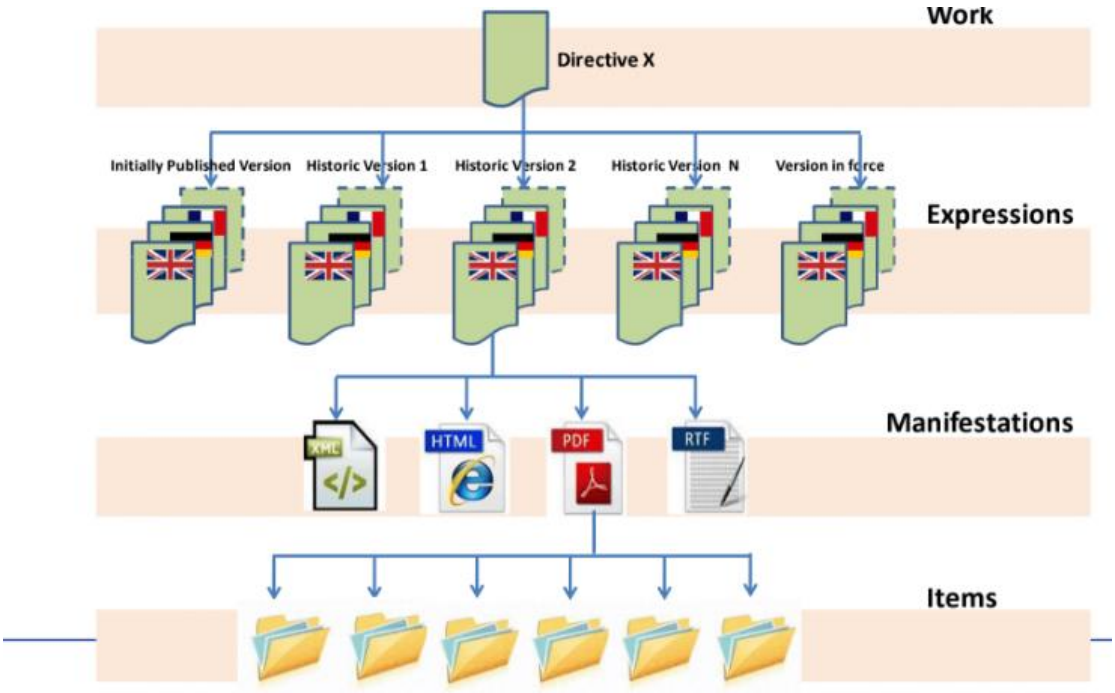
• *Το πρόγραμμα AkomaNtoso* (http://www.akomantoso.org) του Σχεδίου Δράσης για Ηλεκτρονικά Κοινοβούλια των Αφρικανικών Χωρών. Τα σχήματα XML του Akoma Ntoso αποσαφηνίζουν τη δομή και τα σημασιολογικά συστατικά των ψηφιακών

εγγράφων, ώστε να δημιουργηθούν πληροφορίες υψηλής αξίας και να αυξηθεί η αποτελεσματικότητα και η λογοδοσία σε κοινοβουλευτικό, νομοθετικό και δικαστικό πλαίσιο. Το XML σχήμα του Akoma Ntoso αξιοποιήθηκε στο πρόγραμμα EUCases που θα αναφερθεί παρακάτω.

• *Το MetaLex* [*Boer, Hoekstra, Winkels, Van Engers, & Willaert, 2002*], το οποίο αρχικά προτάθηκε ως λεξιλόγιο XML για την κωδικοποίηση της δομής και του περιεχομένου των νομοθετικών εγγράφων, αλλά στην συνέχεια επικαιροποιήθηκε με λειτουργίες σχετικές με τη χρονομέτρηση και τη διαχείριση εκδόσεων των εγγράφων. Μετά την υιοθέτησή του από την Ευρωπαϊκή Επιτροπή Τυποποίησης (CEN), η MetaLex εξελίχθηκε σε οντολογία OWL με την ονομασία CEN MetaLex [*Chalkidis, Nikolaou, Soursos, & Koubarakis, 2017]*.

• *Το «Ευρωπαϊκό Αναγνωριστικό της Νομοθεσίας» - European Legislation Identifier* [εφεξής ELI] που αποτελεί μια ειδική πρωτοβουλία για την παροχή URI HTTP για τον προσδιορισμό της ευρωπαϊκής νομοθεσίας [*Council of European Union,* 2012]. Το πρότυπο ELI συμμορφώνεται με τις οδηγίες των Ανοιχτών Συνδεδεμένων Δεδομένων και παρέχει ένα ελάχιστο σύνολο ομοιόμορφων μεταδεδομένων για την τυποποίηση των νομοθετικών βάσεων δεδομένων και για το πλαίσιο των νομικών εφημερίδων. Η χρήση του προτύπου ELI επιτρέπει: α) την αναγνώριση όχι μόνο κάθε νομοθετικής πράξης αλλά και των διαρθρωτικών της στοιχείων (τμήματα, κεφάλαια , άρθρα, παραρτήματα κλπ), β) τους συνδέσμους σε όλα τα δομικά στοιχεία και γ) τις παραπομπές / συνδέσεις σε ιστορικές εκδόσεις ή σε άλλες γλώσσες. Το πρότυπο ELI χρησιμοποιεί μοναδικά αναγνωριστικά πόρων (URI) και εφαρμόζεται ήδη σε ορισμένα κράτη μέλη της ΕΕ, π.χ. στο Λουξεμβούργο.

Εικόνα 8: Το ELI ακολουθεί το μοντέλο FRBR των βιβλιοθηκών

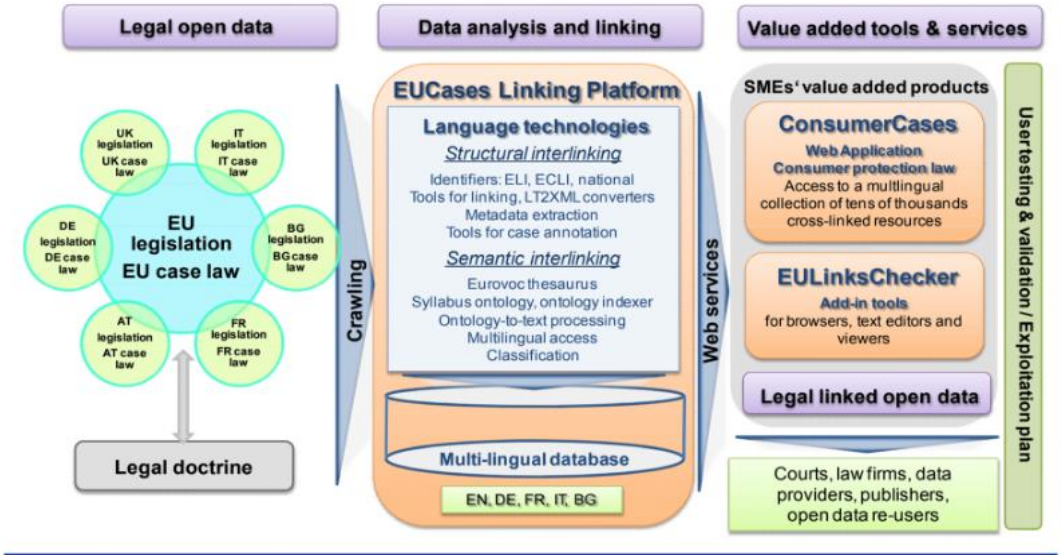


• *Το «Ευρωπαϊκό Αναγνωριστικό της Νομολογίας» - European Case Law Identifier* [εφεξής ECLI] που αναπτύχθηκε για τη διευκόλυνση της ορθής και αδιαμφισβήτητης παραπομπής των δικαστικών αποφάσεων που έχουν εκδοθεί από τα ευρωπαϊκά και εθνικά δικαστήρια και σχετίζονται με το δίκαιο της ΕΕ. Με το ECLI δημιουργήθηκε ένα σύνολο ομοιόμορφων μεταδεδομένων που θα συμβάλει στη βελτίωση των μηχανισμών αναζήτησης της νομολογίας. Το ECLI αποτελείται από πέντε στοιχεία: α) τη συντομογραφία ECLI, β) τον κωδικό χώρας, γ) τη συντομογραφία του δικαστηρίου, δ) το έτος της απόφασης με τέσσερα ψηφία και ε) έναν μοναδικό αύξοντα αριθμό, μέχρι 25 αλφαριθμητικούς χαρακτήρες, σε μορφή που αποφασίζεται από κάθε κράτος μέλος. Τα στοιχεία χωρίζονται από άνω κάτω τελεία. Παράδειγμα ECLI θα μπορούσε να είναι: ECLI: NL: HR: 2009: 384425 που αντιστοιχεί σε Απόφαση 384425 του Ανώτατου Δικαστηρίου («Hoge Raad») των Κάτω Χωρών («NL») του έτους 2009 [*Stroetmann, 2013*].

• *Το πρόγραμμα EUCases* που υλοποιήθηκε με την συμμετοχή 6 επιλεγμένων κρατών – μελών της ΕΕ (Αυστρία, Βουλγαρία, Γερμανία, Γαλλία, Ιταλία και Ηνωμένο Βασίλειο). Για την εκτέλεση του έργου αναπτύχθηκαν τεχνικές και εργαλεία όπως το εργαλείο μετατροπής «Νομικό κείμενο προς XML» (Legal Text 2 XML-εφεξής LT2XML), το οποίο μετασχηματίζει τα έγγραφα που αποκτήθηκαν από τις ανοικτές πύλες δεδομένων της Ευρωπαϊκής Ένωσης (EUR-Lex) και τα παραπάνω έξι κράτη - μέλη της ΕΕ σε «νομική γλώσσα XML» (Legal XML) όπως καθορίζεται από το πρότυπο Akoma Ntoso

για τη διαχείριση εγγράφων. Τα μετασχηματισμένα ανοιχτά νομικά δεδομένα που χρησιμοποιήθηκαν στην πλατφόρμα σύνδεσης EUCases συμμορφώνονται όχι μόνο με το σχήμα Akoma Ntoso, αλλά και με το ελάχιστο σύνολο μεταδεδομένων που ορίζεται από τα ευρωπαϊκά πρότυπα ELI και ECLI. Κατόπιν της συλλογής των ανοιχτών νομικών δεδομένων στη βάση δεδομένων της EuCases, γινόταν η μετατροπή τους σε ομοιογενή μορφή XML με τα εργαλεία LT2XML, το αποτέλεσμα αποθηκευόταν στη βάση δεδομένων και τέλος τα μεταδεδομένα δημοσιεύονταν ως Linked Open Data σε μορφή RDF [*EuCases Project, 2014*] [βλ. εικόνα 9,10].

Εικόνα 9: Επισκόπηση του έργου Eucases



Εικόνα 10: Ο Συνδεδεμένος Ανοικτός κύκλος ζωής δεδομένων του Eucases

Η υλοποίηση του ως άνω προγράμματος είχε ως συνέπεια: α) την ανάπτυξη_των web εφαρμογών ConsumerCases και EuroCases που παρέχουν πρόσβαση σε μια πολύγλωσση συλλογή εθνικών δικαστικών αποφάσεων, επεξεργασμένων σημασιολογικά και διασυνδεδεμένων με την κοινοτική και την εθνική νομοθεσία, β) τη δημιουργία του ανοιχτού συνόλου δεδομένων Eucases και της πλατφόρμας σύνδεσης EUCases που διαθέτει SPARQL endpoint και γ) τη δημιουργία του EULinksChecker, το οποίο αναπτύσσεται ως μια σειρά πρόσθετων εργαλείων - εφαρμογών που μπορούν να εγκατασταθούν (συνδεδεμένες μέσω του διαδικτύου στην πλατφόρμα σύνδεσης EUCases) και να ενσωματωθούν στα πιο δημοφιλή προϊόντα λογισμικού π.χ Microsoft Word και μηχανές αναζήτησης π.χ. Google Chrome [*EuCases Project*, 2015].

• *Το CELLAR* που είναι το αρχείο καταγραφής μεταδεδομένων των εγγράφων (νομικών και μη) που παράγονται από τα ευρωπαϊκά θεσμικά όργανα. Περιλαμβάνει περίπου 150 εκατομμύρια έγγραφα σε 24 γλώσσες και τα μεταδεδομένα αποθηκεύονται και περιγράφονται ως τριπλέτες RDF (περίπου 800 εκατομμύρια στο σύνολο). Το Cellar είναι επίσης η πηγή πληροφοριών της δικτυακής πύλης Eur-Lex (http://www.eurlex.eu) η οποία παρέχει πρόσβαση σε διάφορους τύπους νομικών μέσων, συμπεριλαμβανομένων των συνθηκών, της νομοθεσίας, της νομολογίας και των νομοθετικών προτάσεων. Επί του παρόντος, το SPARQL endpoint του Cellar λαμβάνει περίπου 10 εκατομμύρια αιτήματα ημερησίως (στατιστικές Cellar του Απριλίου 2018) [*Francesconi, 2018*].

• *Η διαδικτυακή πλατφόρμα https://www.legislation.di.uoa.gr/ και η ελληνική εφαρμογή Nomothesia* που δημιουργήθηκε από Έλληνες πανεπιστημιακούς, οι οποίοι ανέπτυξαν μια οντολογία OWL με το όνομα Οντολογία Νομοθεσίας, για τη μοντελοποίηση του περιεχομένου των ελληνικών νομοθετικών εγγράφων, τη δυνατότητα διατύπωσης ερωτημάτων (queries) σχετικά με το περιεχόμενο της ελληνικής νομοθεσίας, παρακολούθησης της εξέλιξης ενός νομοθετικού εγγράφου και αποθήκευσης δεδομένων. Περαιτέρω, η ομάδα υλοποίησης του έργου προχώρησε στην κατασκευή ενός εργαλείου προσαρμοσμένου στις ιδιαιτερότητες της Ελληνικής Νομοθεσίας, που περιλαμβάνει όλα τα νομοθετικά κείμενα που δημοσιεύθηκαν στην Εφημερίδα της Κυβερνήσεως κατά την περίοδο 2006-2015 διαμορφωμένα κατά τρόπο που επιτρέπει τη σύνδεση του προκύπτοντος συνόλου δεδομένων τις ελληνικές διαδικτυακές πύλης ανοιχτών δεδομένων της DBpedia και του http://linkedopendata.gr/ . Έχοντας υπόψη το πρότυπο ELI, δημιουργήθηκε ένα εργαλείο ανάλυσης βασισμένο σε κανόνες που έχει σχεδιαστεί για να είναι ευέλικτο και ανθεκτικό στα συχνότερα σφάλματα κατά τη διαδικασία πληκτρολόγησης ενός νομοθετικού εγγράφου, κοινές αποκλίσεις από την κωδικοποίηση

της ελληνικής νομοθεσίας καθώς και αναγνώριση ορισμένων φράσεων που υποδηλώνουν τροποποιήσεις και καταργήσεις (G3 Parser). Συνολικά, το εν λόγω εργαλείο επεξεργάστηκε 2.676 νομικά έγγραφα και αποθήκευσε περίπου 1.85 εκατομμύρια τριπλέτες RDF. Επιπλέον, η ομάδα το έργου ανέπτυξε μια πρωτοποριακή διαδικτυακή εφαρμογή αποκαλούμενη Nomothesia, η οποία προσφέρει προηγμένες λειτουργίες προβολής, αναζήτησης και ανάλυσης της ελληνικής νομοθεσίας, ενώ διαθέτει παράλληλα ένα SPARQL endpoint και ένα RESTful API που επιτρέπουν τη σύνταξη πολύπλοκων ερωτημάτων. Ως αποτέλεσμα, ένα πολύ σημαντικό μέρος της νομοθεσίας της Ελλάδας συνδέθηκε με εξωτερικές πηγές και δημοσιοποιήθηκε σύμφωνα με τις αρχές του Σημασιολογικού Ίστού για τα ανοιχτά δεδομένα για πρώτη φορά [*Chalkidis, Nikolaou, Soursos, & Koubarakis, 2017*].

• *Το Semantic Finlex:* Από την συγκριτική επισκόπηση της βιβλιογραφίας προέκυψε ότι η πλέον προηγμένη, πλήρης και επικαροποιημένη πλατφόρμα δημοσίευσης, λήψης, επαναχρησιμοποίησης και διασύνδεσης νομικών δεδομένων είναι το «Semantic Finlex» όπου δημοσιεύεται το Φιλανδικό Δίκαιο και τα σχετιζόμενα έγγραφα (όπως δικαστικές αποφάσεις), σύμφωνα με τις αρχές και τις τεχνικές των Ανοιχτών Συνδεδεμένων Δεδομένων. Την εν λόγω πλατφόρμα επιλέξαμε και για την άντληση και ανάλυση των δεδομένων που περιέχεται στο τέταρτο κεφάλαιο της παρούσας εργασίας. Η προσπάθεια δημιουργίας της υπηρεσίας Semantic Finlex αποσκοπεί να παράσχει τη νομοθεσία και τη νομολογία ως Linked Open Data μέσω απλών APIs συνδεδεμένων δεδομένων και να συνδέσει τα σύνολα δεδομένων μεταξύ τους. Η βασική έμπνευση των δημιουργών της νέας αυτής υπηρεσίας ήταν ο διακομιστής εγγράφων MetaLex [http://doc.metalex.eu], ο οποίος περιέχει την Ολλανδική νομοθεσία ως Ανοιχτά Συνδεδεμένα Δεδομένα χρησιμοποιώντας την γλώσσα CEN Metalex XML και τις πρότυπες οντολογίες [*Oksanen, et al., 2017*]. Στην σημερινή της μορφή η Semantic Finlex αποτελείται από τέσσερα διαφορετικά σύνολα δεδομένων: α) περίπου 47.000 νομοθετικά έγγραφα, β) περίπου 2.800 άρθρα, γ) περίπου 5.500 αποφάσεις του Ανωτάτου Δικαστηρίου από το 1980 μέχρι σήμερα και δ) περίπου 9.300 αποφάσεις του Ανώτατου Διοικητικού Δικαστηρίου από το 1987 μέχρι σήμερα. Όλα τα παραπάνω σύνολα δεδομένων μετασχηματίζονται από διαφορετικές μορφές παλαιού τύπου σε RDF, προσαρμοσμένα στα μοντέλα δεδομένων που προηγουμένως αναφέρθηκαν. Νέα και ενημερωμένα έγγραφα αποστέλλονται κάθε εβδομάδα από την υπηρεσία Finlex και μετατρέπονται σε RDF [https://data.finlex.fi/en/main] [*Oksanen et al.,2018, Frosterus, Tuominen & Hyvonen, 2014*].

Τα παραπάνω νομικά συνδεδεμένα δεδομένα του Semantic Finlex δημοσιεύονται στην πλατφόρμα LDF.fi της Φινλανδίας, η οποία υιοθετεί την άποψη ότι για να είναι δυνατή η ουσιαστική χρήση των συνόλων δεδομένων, χρειάζεται μεγαλύτερη υποστήριξη των καταναλωτών, καλύτερη πρόσβαση στα δεδομένα και πληρέστερη κατανόηση της ποιότητας και προέλευσής τους. Για το σκοπό αυτό, η πλατφόρμα LDF.fi διευρύνει το μοντέλο των πέντε αστέρων του Tim Berners-Lee με δύο πρόσθετα αστέρια:

★★★★★★ Παροχή των δεδομένων με ένα σαφές σχήμα και τεκμηρίωση, έτσι ώστε οι άνθρωποι να μπορούν εύκολα να κατανοούν και να επαναχρησιμοποιούν τα δεδομένα και τα μεταδεδομένα.
★★★★★★★ Επαλήθευση των δεδομένων και δήλωση της προέλευσής τους έτσι ώστε οι άνθρωποι να μπορούν να εμπιστεύονται την ποιότητά τους.

Ακολουθώντας τις προαναφερθείσες τεχνικές, το Semantic Finlex δημοσιεύει πλέον τα δεδομένα σε μορφή που συμμορφώνεται με το πρότυπο RDF του Σημασιολογικού Ιστού. Επιπλέον, η υπηρεσία παρέχει τα πρωτότυπα έγγραφα Finlex ως τράπεζες δεδομένων σε μορφή XML, περιλαμβάνοντας πρωτότυπα κείμενα πράξεων και διαταγμάτων, καθώς και νομικές υποθέσεις. Οι βασικές λειτουργίες του Semantic Finlex περιλαμβάνουν: α) τη δυνατότητα λήψης μέρους ή του συνόλου των πληροφοριών σε διάφορες μορφές (RDF Turtle, JSON κλπ.) από την εφαρμογή του λήπτη, β) την ανάκτηση των πληροφοριών με βάση το URI και πιο συγκεκριμένα με την πληκτρολόγηση του URI στο πρόγραμμα περιήγησης ή μέσω της προγραμματιστικής ανάκτησης δεδομένων (ως πηγή αναγνωριστικών, το Semantic Finlex χρησιμοποιεί το σύστημα ELI) και γ) την ανάκτηση δεδομένων μέσω της διασύνδεσης SPARQL σε εφαρμογές ή μόνο για προβολή.

2.4 Η ΕΞΟΡΥΞΗ ΚΑΙ ΑΝΑΛΥΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ (Data Mining and Analysis)

2.4.1 Έννοια και στόχος της εξόρυξης δεδομένων

Αφού μελετήθηκαν και τα ανοιχτά συνδεδεμένα νομικά δεδομένα, η βιβλιογραφική έρευνα στράφηκε στην επιστήμη της εξόρυξης και ανάλυσης των δεδομένων. Η εξόρυξη δεδομένων (data mining), συχνά αποκαλούμενη και ως ανακάλυψη γνώσεων από βάσεις δεδομένων [Knowledge Discovery in Databases-KDD], ορίζεται ως η εξεύρεση μια ενδιαφέρουσας, μη προφανούς και πιθανόν χρήσιμης πληροφορίας ή προτύπων από

μεγάλες βάσεις δεδομένων, με την χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν κατανοητή δομή και να βοηθήσουν στην λήψη των κατάλληλων αποφάσεων [*Witten, Eibe & Hall, 2011*].

2.4.2 Η Διαδικασία εξόρυξης των δεδομένων

Η εξόρυξη δεδομένων περιλαμβάνει ολόκληρη τη διαδικασία από την συλλογή των δεδομένων μέχρι την προβολή και την εφαρμογή πρότυπων ευρημάτων σε νέες, άγνωστες δομές δεδομένων. Η διαδικασία περιλαμβάνει: (α) τεχνικές για την προεπεξεργασία δεδομένων, (β) το πραγματικό σύστημα εξόρυξης δεδομένων (σύστημα DM), (γ) την οπτικοποίηση και επικύρωση των δεδομένων και (δ) την ερμηνεία και την αξιολόγηση των δεδομένων που οδηγεί στη γνώση.

Τα βήματα του data mining είναι συνοπτικά [εικόνα 11]:

• *Η Ενσωμάτωση των δεδομένων (data integration):* αρχικά συλλέγονται τα δεδομένα από διάφορες πηγές (Excel, MS Access, Oracle, SQL Server, csv, rdf stores, XML κ.λπ.) σε μια μοναδική πηγή δεδομένων που ονομάζεται target data/database με τη χρήση κάποιας τεχνολογίας όπως SPARQL, SQL, PYTHON κλπ.

• *Η Επιλογή των δεδομένων (data selection):* σε αυτό το βήμα επικεντρωνόμαστε μόνο σε εκείνα τα δεδομένα που απαιτούνται για την εκπλήρωση της έρευνας / την παραδοχή. Αυτό πρακτικά σημαίνει ότι επιλέγονται μόνο τα σημαντικά δεδομένα που είναι αναγκαία για να προχωρήσουμε.

• *Ο Καθαρισμός και η κανονικοποίηση των δεδομένων (data cleansing and normalization):* τα δεδομένα που εισάγονται από τις διάφορες πηγές μπορεί να έχουν διαφορετική μορφή από την target database. Για τον λόγο αυτό πρέπει να καθαρίσουμε τα δεδομένα χρησιμοποιώντας τον κατάλληλο αλγόριθμο.

• *Ο Μετασχηματισμός των δεδομένων (data transformation):* στην συνέχεια, τα δεδομένα προετοιμάζονται και μετατρέπονται σε τυπική μορφή.

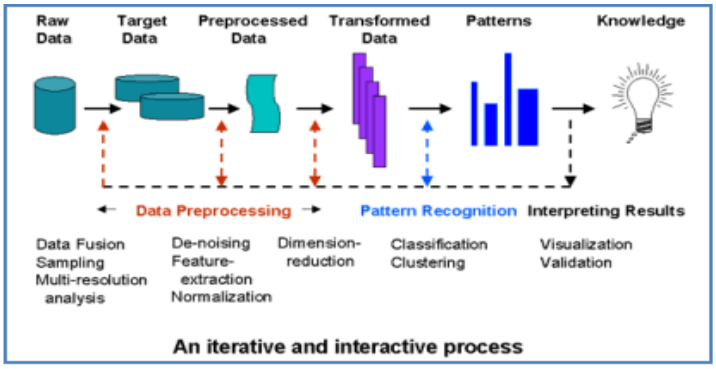
• *Η Εξόρυξη των δεδομένων (data mining):* στο επόμενο βήμα αναλύουμε και προσδιορίζουμε τον τύπο του αλγόριθμου εξόρυξης δεδομένων που είναι κατάλληλος για τα δεδομένα που συλλέχθηκαν και στη συνέχεια εφαρμόζουμε αλγόριθμους για τον

προσδιορισμό των κρυφών προτύπων. Παραδείγματος χάριν με την χρήση του k-means clustering algorithm κατηγοριοποιούνται τα δεδομένα σε ομάδες ανάλογα με τη συσχέτιση τους, χωρίς να είναι εκ των προτέρων γνωστές αυτές οι συσχετίσεις, με την τεχνική του correlation αναλύεται η εγγύτητα της συσχέτισης μεταξύ δύο ή περισσοτέρων μεταβλητών κλπ.

• *Η Αξιολόγηση μοτίβου (pattern evaluation):* το πρότυπο που εντοπίσαμε από τα δεδομένα ερμηνεύεται και αξιολογείται στη συνέχεια για να αποκτηθούν γνώσεις από αυτό.

• *Η Παρουσίαση γνώσης (knowledge presentation):* Αυτός είναι ο στόχος της τεχνικής εξόρυξης δεδομένων, όπου η γνώση που συλλέγεται από την παραπάνω διαδικασία οπτικοποιείται, αξιολογείται και λαμβάνεται υπόψη για την λήψη κρίσιμων αποφάσεων.

Εικόνα 11: Η διαδικασία του data mining



Εικόνα 12: Η συμβολή διαφορετικών αρχών στη διαδικασία του data mining.

2.4.3 Οι Τύποι της εξόρυξης δεδομένων

Οι Τύποι της εξόρυξης δεδομένων κατηγοριοποιούνται ως εξής:

• *Εξόρυξη δεδομένων (data mining),* η οποία περιλαμβάνει την ανάλυση αριθμητικών και απόλυτων δεδομένων που είναι αποθηκευμένα σε μεγάλα και πολύπλοκα σύνολα δεδομένων. Συχνά ο όρος αυτός χρησιμοποιείται για να περιγράψει πιο εξειδικευμένες τεχνικές, όπως η εξόρυξη κειμένου, ιστού ή χώρου.

• *Εξόρυξη κειμένου (text mining)*: ο συγκεκριμένος τύπος εξόρυξης περιλαμβάνει αλγορίθμους για την ανάλυση λεξικών και γραμματικών πτυχών των κειμένων. Το πραγματικό κείμενο αναλύεται σε συγκεκριμένες δομές όπου τα πρότυπα και οι βασικές πληροφορίες του κειμένου καταγράφονται, ομαδοποιούνται και ταξινομούνται χρησιμοποιώντας τις μεθόδους εξόρυξης δεδομένων. Τα πρώτα εργαλεία εξόρυξης κειμένου μπορούσαν να καταγράφουν τα περιεχόμενα και τις δομές απλών εγγράφων κειμένου όπως τα έγγραφα του Microsoft Word και του Acrobat PDF, ενώ πλέον έχουν τη δυνατότητα να σαρώνουν και να αναλύουν το αδόμητο κείμενο σε s, memos, έρευνες, συνομιλίες, σημειώσεις, φόρουμ και παρουσιάσεις [*Begum, 2013*].

• *Εξόρυξη Ιστού (web mining)* που συνιστά την εφαρμογή μεθόδων εξόρυξης δεδομένων σε πληροφορίες που συλλέγονται στο Διαδίκτυο. Στην εξόρυξη ιστού, γίνεται διάκριση μεταξύ (1) εξόρυξης περιεχομένου ιστού, η οποία είναι η ανάλυση του περιεχομένου του ιστότοπου, (2) εξόρυξη δομής διαδικτύου ή σχέσεων, δηλαδή ανάλυση εισερχόμενων και εξερχόμενων υπερσυνδέσμων ιστοτόπων και (3) εξόρυξη χρήσης ιστού, η οποία καταγράφει και αναλύει την αλληλεπίδραση των χρηστών με τους ιστότοπους μέσω της σάρωσης αρχείων καταγραφής.

• *Εξόρυξη εικόνας (image mining):* ο στόχος των τεχνικών εξόρυξης εικόνας είναι η ανάλυση και εξαγωγή χωρικών μοτίβων σε δεδομένα εικόνας τα οποία δεν αποθηκεύονται ρητά στις εικόνες. Η εξαγωγή των μοτίβων γίνεται με διάφορους τρόπους όπως π.χ με την αναγνώριση της ύπαρξης και κατανομής των χρωμάτων, της υφής, του σχήματος, των αποστάσεων και των εντάσεων στα δεδομένα εικόνας.

• *Εξόρυξη ζωγραφιών, δεδομένων βίντεο και μουσικής (picture, video data and music mining)*: οι εν λόγω τεχνικές εξόρυξης χρησιμοποιούνται όλο και περισσότερο για να αναγνωρίζουν χαρακτηριστικά σε εικόνες, βίντεο και μουσικά δεδομένα. Αυτός ο τύπος εξόρυξης είναι ο μόνος που μπορεί να αντιμετωπίσει τις τεράστιες ποσότητες περίπλοκων δεδομένων που δημιουργούνται, για παράδειγμα, στο Google ή στο YouTube. Ιδιαίτερη

σημασία έχει η γρήγορη ενεργοποίηση των περιεχομένων ανάκτησης εικόνων / βίντεο, η ευρετηρίαση, ταξινόμηση και παρακολούθηση.

• *Εξόρυξη δεδομένων χρονοσειράς (time series data mining)*: σε αυτό το σύστημα εξόρυξης δεδομένων, οι χρονικές σχέσεις διερευνώνται μέσω μιας ειδικής λειτουργίας απόστασης όπως η δυναμική χρονική στρέβλωση. Ο στόχος είναι να αναγνωριστούν ομοιότητες κατά τη διάρκεια της χρονοσειράς, ακόμα και όταν τα παρόμοια χαρακτηριστικά μετατοπίζονται κατά την πορεία της διαδικασίας. Ένα παράδειγμα είναι το "Google Correlate" (http://www.google.com/trends/ correlate).

• *Χωρική εξόρυξη δεδομένων (spatial data mining)* που επιδιώκει την ανακάλυψη μοτίβων σε μεγάλα, πολυδιάστατα σύνολα χωρικών δεδομένων, τα οποία δημιουργούνται με τεχνικές τηλεπισκόπησης κατά την παρατήρηση της Γης. Εφόσον στην ανάλυση μοτίβων ενσωματωθούν, πέραν των χωρικών δεδομένων, και πρόσθετες χρονολογικές σειρές, τότε χρησιμοποιείται ο όρος εξόρυξη χωρο-χρονικών δεδομένων *[Lausch, Schmidt & Tischendorf, 2015]*.

2.4.4 Τα μοντέλα ανάλυσης των δεδομένων

Τα τέσσερα βασικά μοντέλα ανάλυσης των δεδομένων που σχετίζονται με τη διαδικασία του data mining είναι τα εξής [*Bekker, 2017*] (βλ. εικόνα 13):

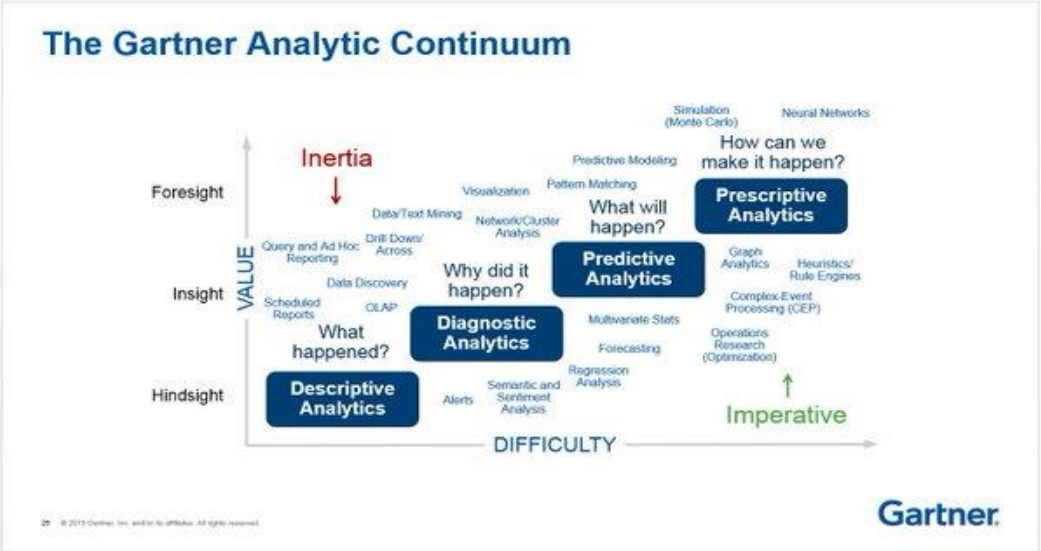
• *Μοντέλο Περιγραφικής Ανάλυσης* (**Descriptive Analytics**)**:** η περιγραφική ανάλυση ζευγαρώνει ακατέργαστα δεδομένα από πολλαπλές πηγές, προκειμένου να δώσει πολύτιμες πληροφορίες σχετικά με το παρελθόν. Η συντριπτική πλειοψηφία των στατιστικών που χρησιμοποιούνται εμπίπτουν σε αυτήν την κατηγορία. Με την χρήση αυτού του μοντέλου δίνεται απάντηση σχετικά με το ***τι έχει συμβεί*** σε οποιαδήποτε χρονική στιγμή και έτσι οι αναλυτές κατανοούν πώς οι συμπεριφορές του παρελθόντος μπορούν να επηρεάσουν τα μελλοντικά αποτελέσματα. Ωστόσο, τα ευρήματα που εξάγονται με την συγκεκριμένη μέθοδο δείχνουν απλά εάν κάτι είναι λάθος ή σωστό, χωρίς να εξηγούν το γιατί. Για το λόγο αυτό, το descriptive model συνδυάζεται συνήθως και άλλους τύπους εξόρυξης και ανάλυσης δεδομένων.

• *Μοντέλο Διαγνωστικής Ανάλυσης* (**Diagnostic Analytics**)**:** με τη χρήση αυτής της μεθόδου ανάλυσης, τα ιστορικά δεδομένα μπορούν να μετρηθούν με άλλα δεδομένα για να απαντηθεί το ερώτημα ***γιατί συνέβη κάτι***. Η διαγνωστική ανάλυση δίνει τη δυνατότητα ανίχνευσης εξαρτήσεων και ταυτοποίησης προτύπων, παρέχοντας τελικώς μια βαθιά γνώση ενός συγκεκριμένου προβλήματος.

• *Μοντέλο Προγνωστικής Ανάλυσης (**Predictive Analytics**):* οι προγνωστικές αναλύσεις χρησιμοποιούν τα ευρήματα των παραπάνω περιγραφικών και διαγνωστικών αναλύσεων για την ανίχνευση τάσεων, συστάδων και εξαιρέσεων και για την κατανόηση και πρόβλεψη του μέλλοντος. Με το μοντέλο αυτό απαντάται το ερώτημα ***τι είναι πιθανό να συμβεί.*** Παρά τα πολυάριθμα πλεονεκτήματα που προσφέρει η προγνωστική ανάλυση, πρέπει να γίνεται κατανοητό ότι η πρόβλεψη είναι απλώς μια εκτίμηση του μέλλοντος με βάσει τις πιθανότητες και όχι βεβαιότητα. Η ακρίβεια της πρόβλεψης εξαρτάται σε μεγάλο βαθμό από την ποιότητα των δεδομένων και τη σταθερότητα της κατάστασης, συνεπώς απαιτεί προσεκτική επεξεργασία και συνεχή βελτιστοποίηση.

• *Μοντέλο Κανονιστικής Ανάλυσης* (**Prescriptive Analytics**): στο εν λόγω μοντέλο ανάλυσης χρησιμοποιούνται εξελιγμένα εργαλεία και τεχνολογίες για να απαντηθεί το ερώτημα ***πώς μπορούμε να κάνουμε κάτι να συμβεί***. Η κανονιστική ανάλυση επωφελείται από τα αποτελέσματα των περιγραφικών και προγνωστικών αναλύσεων για να υποδείξει τις διαθέσιμες επιλογές σχετικά με τον τρόπο αξιοποίησης μιας μελλοντικής ευκαιρίας ή την αποτροπή ή ελαχιστοποίηση ενός μελλοντικού κινδύνου και δείχνει την επίδραση κάθε απόφασης. Με αυτή τη μέθοδο λαμβάνονται συνεχώς νέα δεδομένα για να επαναπροσδιορίζουν και βελτιώνουν την ακρίβεια πρόβλεψης και να κατευθύνουν στην επιλογή των βέλτιστων στρατηγικών αποφάσεων.

Εικόνα 13: Τα μοντέλα ανάλυσης [Gartner, 2015]



2.4.5. Οι βασικές μέθοδοι εξόρυξης δεδομένων

Από τη μελέτη της βιβλιογραφίας προέκυψε ότι υπάρχουν διάφορες μέθοδοι υποστήριξης της εξόρυξης δεδομένων [εικόνα 14], οι οποίες μπορούν να ομαδοποιηθούν ανάλογα με τον στόχο που επιδιώκουν. Μια επισκόπηση των πιο κοινών μεθόδων εξόρυξης δεδομένων είναι η ακόλουθη.

• *Σύνδεση και ανάλυση αλληλουχίας (**association and sequence analysis**):* Με τη βοήθεια της συγκεκριμένης μεθόδου, οι σχέσεις μεταξύ των αντικειμένων μπορούν να υλοποιηθούν και να ποσοτικοποιηθούν. Χρησιμοποιώντας συγκεκριμένους δείκτες, οι οποίοι στις περισσότερες περιπτώσεις περιλαμβάνουν την υποστήριξη, την εμπιστοσύνη και τις τιμές ανύψωσης, η ισχύς των αναγνωρισμένων ενώσεων μπορεί να αξιολογηθεί.

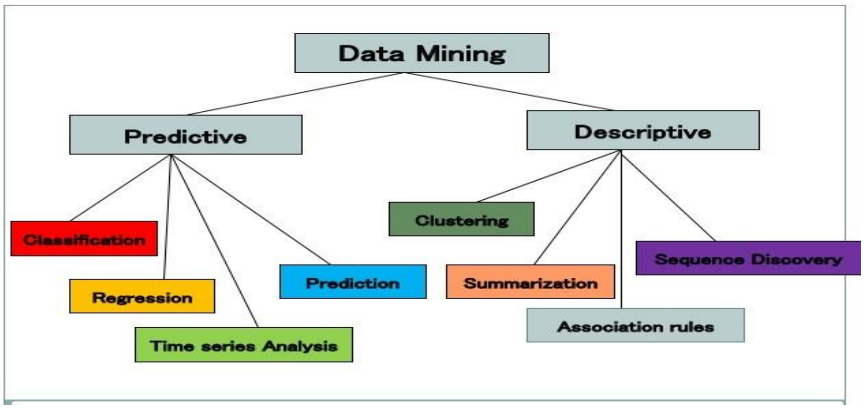
• *Ομαδοποίηση και συσπείρωση (**grouping and clustering**)*: με τις συγκεκριμένες διαδικασίες συγκεντρώνονται παρόμοια αντικείμενα σε ομάδες, έτσι ώστε τα πολύ παρόμοια αντικείμενα να συμπεριληφθούν μέσα σε μία ομάδα και οι ομάδες να διαφέρουν όσο το δυνατόν περισσότερο μεταξύ τους. Για τον χαρακτηρισμό των ομοιοτήτων των ομάδων χρησιμοποιούνται πολυάριθμες μετρήσεις.

• *Παλινδρόμηση (**Regression**):* Με τη βοήθεια της ανάλυσης παλινδρόμησης, καθορίζονται οι λειτουργικές εξαρτήσεις μεταξύ των μεταβλητών που περιλαμβάνονται μέσα σε ένα σύνολο δεδομένων. Τα μοντέλα παλινδρόμησης χρησιμοποιούνται για την εκτίμηση ή την πρόβλεψη των μεταβλητών. Για να αντιπροσωπεύσουμε εξαρτήσεις υπάρχουν γραμμικές (linear) και μη γραμμικές (non-linear) προσεγγίσεις (π.χ. τετραγωνικές, λογικές ή Poisson) [*Witten, Eibe & Hall, 2011*].

• *Ταξινόμηση (**classification**)*: σκοπός της ταξινόμησης είναι η εύρεση λειτουργιών και μοντέλων με τη βοήθεια των οποίων τα αντικείμενα δεδομένων μπορούν να ανατεθούν σε ήδη αναγνωρισμένες κλάσεις. Με τη βοήθεια ενός προκαθορισμένου μοντέλου ταξινομούνται αντικείμενα που δεν έχουν καμία γνωστή ταξινόμηση. Τα μοντέλα μπορούν να προσδιοριστούν με τη βοήθεια νευρωνικών δικτύων, διακριτικών αναλύσεων ή decision trees και τυχαίων δασών.

Υπάρχουν και διάφορες άλλες μέθοδοι όπως η ανάλυση χρονικών σειρών ή η απεικόνιση και οι εξελικτικοί αλγόριθμοι.

Εικόνα 14: Μέθοδοι εξόρυξης δεδομένων που χρησιμοποιούν τα μοντέλα προγνωστικής και περιγραφικής ανάλυσης



2.4.6 Η επιρροή των ανοιχτών συνδεδεμένων δεδομένων στην εξέλιξη του data mining

Όπως αναφέρθηκε και παραπάνω, ο στόχος των Ανοιχτών Συνδεδεμένων Δεδομένων είναι η δημοσίευση αλληλένδετων συνόλων δεδομένων με τη χρήση ερμηνευτικής σημασιολογίας μηχανών. Η ανάπτυξη των Ανοιχτών Συνδεδεμένων Δεδομένων, όπως είναι φυσικό, οδήγησε σε νέες συσχετίσεις και επηρέασε την επιστήμη του data mining. Κι αυτό διότι μέσω των συνδέσμων που δημιουργούνται μεταξύ των δεδομένων στο Σημασιολογικό Ιστό έχουμε τη δυνατότητα να ανακαλύψουμε εντελώς νέες σχέσεις, πρότυπα και γνώσεις. Επίσης, η σύζευξη των Ανοιχτών Συνδεδεμένων Δεδομένων τυποποιείται μέσω κανόνων RDF και σύνταξης και συνεπώς δεν απαιτούνται περαιτέρω βήματα για τον μετασχηματισμό ή την ενσωμάτωση των δεδομένων. Κατά συνέπεια, η εξόρυξη και ανάλυση των πληροφοριών μέσω των κατάλληλων εργαλείων που υποστηρίζουν ερωτήματα σε Linked Open Data (όπως πχ το LOD extension του RapidMiner https://dws.informatik.uni-mannheim.de/ en/research/rapidminer lod extension, παρακάτω υπό 3.5), διευκολύνεται σημαντικά. Υπάρχουν δύο βασικές στρατηγικές χρήσης των Ανοιχτών Συνδεδεμένων Δεδομένων για εξόρυξη των δεδομένων: (α) ανάπτυξη εξειδικευμένων μεθόδων εξόρυξης για Linked Open Data και (β) προεπεξεργασία των Linked Open Data έτσι ώστε να μπορούν να προσεγγιστούν με παραδοσιακές μεθόδους εξόρυξης δεδομένων.

Οι αναλυτές αλλά και οι απλοί χρήστες, αφού λάβουν ένα τοπικό σύνολο δεδομένων (όπως μια σχεσιακή βάση δεδομένων), το συνδέουν με τις αντίστοιχες έννοιες των Ανοιχτών Συνδεδεμένων Δεδομένων από το επιλεγμένο dataset. Μόλις οριστούν οι σύνδεσμοι (links), μπορούν να διερευνηθούν οι εξερχόμενες συνδέσεις σε εξωτερικά σύνολα Ανοιχτών Συνδεδεμένων Δεδομένων. Στην συνέχεια, εφαρμόζονται οι

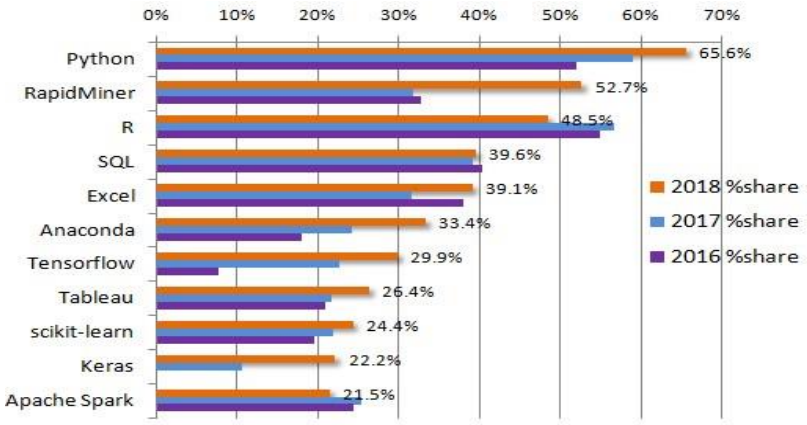
κατάλληλες τεχνικές για την ενοποίηση (consolidation) και τον καθαρισμό (cleansing) των δεδομένων και στο επόμενο βήμα εκτελείται μετασχηματισμός των δεδομένων που έχουν συλλεχθεί έτσι ώστε τα δεδομένα να γίνουν επεξεργάσιμα από οποιοδήποτε αλγόριθμο ανάλυσης δεδομένων. Κατόπιν εφαρμόζεται η κατάλληλη μέθοδος εξόρυξης των δεδομένων και στο τέλος παρουσιάζονται τα αποτελέσματα της παραπάνω διαδικασίας [*Ristoski, 2015*].

Εικόνα 15: Διαδικασία εξόρυξης ανοιχτών συνδεδεμένων δεδομένων

2.4.7 Τα Εργαλεία εξόρυξης των δεδομένων

Τα τελευταία χρόνια, τα εργαλεία εξόρυξης δεδομένων έχουν γίνει πολύ πιο διαθέσιμα, διαδραστικά και διερευνητικά. Αυτή η εξέλιξη έχει δώσει τη δυνατότητα σε άτομα που προέρχονται από διαφορετικούς κλάδους και δεν διαθέτουν απαραίτητα εξειδικευμένη εμπειρία σε προγραμματισμό υπολογιστών να διεξάγουν εξόρυξη δεδομένων στο πεδίο έρευνάς τους. Κατά τη δημοσκόπηση της KDnuggets [https://www.kdnuggets.com/2018/05/poll-tools-analytics- data- science -machine-learning-results.html] (βλ. εικόνα 16) η χρήση εργαλείων ανοιχτού κώδικα και εμπορικών εργαλείων ανάλυσης, εξόρυξης δεδομένων και ηλεκτρονικής μάθησης τα έτη 2016-2018 διαμορφώθηκε ως εξής:

Εικόνα 16: Αποτελέσματα της δημοσκόπησης KDnuggets 2016-2018:



Σύμφωνα με τα αποτελέσματα της παραπάνω δημοσκόπησης, στην οποία συμμετείχαν 2.052 άτομα, οι γλώσσες προγραμματισμού **Python** και **R** και το εργαλείο εξόρυξης δεδομένων ανοιχτού κώδικα **RapidMiner** χρησιμοποιούνται πιο συχνά από τους χρήστες. Εμείς επιλέξαμε το RapidMiner *(http://www.rapidminer.com)*, το οποίο αποτελεί μια πλατφόρμα ανοιχτού κώδικα που υποστηρίζει διάφορους τύπους εξόρυξης δεδομένων, όπως η εξόρυξη δεδομένων κειμένου, διαδικτύου, εικόνας ή συνδεδεμένων ανοιχτών δεδομένων. Μέσω της εξελιγμένης γραφικής διεπαφής χρήστη, οι διεργασίες εξόρυξης δεδομένων μπορούν να υλοποιηθούν και να εκτελεστούν γρήγορα, διαισθητικά και χωρίς οι χρήστες να διαθέτουν εξειδικευμένες γνώσεις προγραμματισμού Η/Υ. Επιπλέον, το RapidMiner προσφέρει πολλά στάδια πριν και μετά την επεξεργασία των δεδομένων που υποστηρίζουν την ολοκλήρωση, εξαγωγή, μετασχηματισμό και φόρτωση των δεδομένων, καθώς και την ανάλυση και αναφορά των πορισμάτων σε ένα ενιαίο και συνεκτικό γραφικό περιβάλλον χρήστη. Επίσης, το RapidMiner προσφέρει μια ολοκληρωμένη επιλογή δραστηριοτήτων εξόρυξης δεδομένων καθώς και μια ενοποίηση των βιβλιοθηκών WEKA και R. Το RapidMiner υλοποιείται σε Java και μπορεί να ενσωματώσει τις διαδικασίες που αναπτύσσονται από τους χρήστες ως plug-ins. Προσφέρει δε μια διεπαφή προγραμματισμού εφαρμογών (Application Program Interface), μέσω της οποίας μπορεί να χρησιμοποιηθεί σε άλλα προγράμματα ως βιβλιοθήκη Java [*Hirudkgar & Sherekar, 2013, Hofmann & Klinkenberg, 2013*].

Η ροή εργασιών (workflow) του RapidMiner συνοψίζεται σε : (α) εισαγωγή του αρχικού πίνακα δεδομένων, (β) σύνδεση με άλλα δεδομένα, (γ) δημιουργία εμπλουτισμένου πίνακα δεδομένων και (δ) ανάλυση των δεδομένων. Περαιτέρω όμως έχουν αναπτυχθεί και επεκτάσεις (extensions) που είναι διαθέσιμες στο https://marketplace.rapidminer.com και προσθέτουν νέες λειτουργίες στα προϊόντα RapidMiner, όπως εξόρυξη κειμένου, ανίχνευση ιστού ή ενσωμάτωση με τα R και Weka.

2.5 ΣΥΜΠΕΡΑΣΜΑ ΒΙΒΛΙΟΓΡΑΦΙΚΗΣ ΑΝΑΣΚΟΠΗΣΗΣ

Από την βιβλιογραφική έρευνα στο πλαίσιο της παρούσας εργασίας προέκυψε ότι η ελεύθερη πρόσβαση, λήψη και επεξεργασία ιδίως των νομικών δεδομένων παρουσιάζει αρκετές δυσχέρειες, κυρίως διότι στις περισσότερες περιπτώσεις οι παρεχόμενες πληροφορίες είναι σε ακατέργαστη μορφή, γεγονός που εμποδίζει το ευρύ κοινό να τις κατανοήσει και να τις εκμεταλλευτεί προς όφελός του. Είναι όμως σαφές ότι η ελεύθερη πρόσβαση και η διασύνδεση μεταξύ παραπάνω κατηγοριών πληροφοριών σύμφωνα με

την μεθοδολογία του Σημασιολογικού Ιστού για τα ανοιχτά συνδεδεμένα δεδομένα αποτελεί προτεραιότητα των κυβερνήσεων σε παγκόσμιο επίπεδο. Η απελευθέρωση και διαλειτουργικότητα των νομικών εγγράφων έχει άμεση επίπτωση και στην επιστήμη της ανάλυσης δεδομένων, από άποψη ερμηνείας των νομικών δεδομένων, έρευνας των αμοιβαίων εξαρτήσεων μεταξύ των νόμων διαφορετικών κρατών και παρακολούθησης της διαχρονικής εξέλιξης της νομοθεσίας, ενώ γίνεται δυνατή η προσπέλαση, αποθήκευση και ανάλυση των δεδομένων όχι μόνο από ειδικούς που διαθέτουν εξειδικευμένες τεχνικές γνώσεις αλλά και από το απλό κοινό, με τη χρήση των εργαλείων εξόρυξης δεδομένων, ακολουθώντας τις μεθόδους ανάλυσης που προεκτέθηκαν και με τις τεχνικές και τη μεθοδολογία που παρατίθεται στο αμέσως επόμενο κεφάλαιο.

3. ΜΕΘΟΔΟΛΟΓΙΑ

3.1 Εισαγωγή

Στο παρόν κεφάλαιο θα περιγραφεί η μεθοδολογία που ακολουθήθηκε ώστε να επιλέξουμε τα κατάλληλα datasets και να αντλήσουμε τα δεδομένα προκειμένου στην συνέχεια να τα αναλύσουμε και να εξάγουμε συμπεράσματα.

3.2. Τα βήματα που ακολουθήθηκαν κατά την εκπόνηση της εργασίας

Το πρώτο βήμα για την επίτευξη του στόχου της εργασίας ήταν να κατανοήσουμε και να εξοικειωθούμε με τις βασικές έννοιες των ανοιχτών συνδεδεμένων δεδομένων και τις τεχνικές του Σημασιολογικού Ιστού μέσα από τη βιβλιογραφία που αναλύθηκε στο προηγούμενο κεφάλαιο. Στην συνέχεια, αρχίσαμε την περιήγηση στο LOD cloud αναζητώντας τα κατάλληλα σύνολα δεδομένων που θα χρησιμοποιούνταν στο κεφάλαιο της ανάλυσης, την περιγραφή τους, την προέλευσή τους και τις συνδέσεις τους με άλλα σύνολα δεδομένων ώστε να εμβαθύνουμε στις τεχνικές και τα πρότυπα. Ακόμη, μελετήσαμε εκπαιδευτικό υλικό από το Τμήμα Εφαρμοσμένης Πληροφορικής – Τεχνολογίες Ιστού και Ανάλυσης Δεδομένων Ιστού του Πανεπιστημίου Μακεδονίας (ΤΙΑΔΙ) σχετικά με την εφαρμογή των ανοιχτών συνδεδεμένων δεδομένων σε applications. Ακολούθως, εξετάσαμε τα διάφορα εργαλεία εξόρυξης και απεικόνισης των δεδομένων, καταλήγοντας στην επιλογή και χρήση του RapidMiner.

3.3 Εκμάθηση της γλώσσας ερωτημάτων SPARQL

Ένας από τους βασικούς μας προβληματισμούς ήταν με ποια μέθοδο θα διασχίσουμε τους RDF γράφους των Linked Open Data και θα αντλήσουμε τα προς ανάλυση δεδομένα για την εργασία μας. Παρατηρήθηκε ότι στο σύνολο της βιβλιογραφίας η άντληση των ανοιχτών συνδεδεμένων δεδομένων πραγματοποιείται μέσω endpoints στα οποία εκτελούνται SPARQL ερωτήματα. Η γλώσσα SPARQL είναι για τα RDF ό, τι ακριβώς είναι η γλώσσα SQL για τις σχεσιακές βάσεις δεδομένων (RDMS). Στην συγκεκριμένη γλώσσα αρκούν βασικές γνώσεις SQL για την σύνταξη και εκτέλεση ερωτημάτων. Για να εξοικειωθούμε με την SPARQL μελετήσαμε την τεκμηρίωση της στο https://www.w3.org/TR/rdf-sparql-query/ και την εφαρμογή της σε διάφορα παραδείγματα στο https://www.w3.org/2001/sw/DataAccess/rq23/examples.html. Στην SPARQL μπορούμε εμείς οι χρήστες να επιλέξουμε τις μεταβλητές των ανοιχτών συνδεδεμένων δεδομένων που μας ενδιαφέρουν σε κάθε περίπτωση, ώστε να

προσδώσουμε σημασία και κατόπιν της σχετικής ανάλυσης να εξάγουμε χρήσιμα συμπεράσματα.

3.4 Μελέτη οντολογιών και επιλογή των προς ανάλυση δεδομένων

Κατά τη διάρκεια συγγραφής της παρούσας εργασίας παρατηρήθηκαν δυσλειτουργίες στο LOD cloud στο οποίο αρχικά περιηγούμασταν για να βρούμε ενδιαφέροντα σύνολα δεδομένων. Ειδικότερα, πολλά endpoints και συνδέσεις μεταξύ των φυσαλίδων που απεικονίζουν τα σύνολα δεδομένα εμφάνιζαν σφάλματα (connection – internal errors), συνεπώς δεν επέτρεπαν την εκτέλεση ερωτημάτων και δυσχέραιναν την ανεύρεση των κατάλληλων για την μελέτη μας datasets. Με αφετηρία το άρθρο του Enrico Francesconi [2018], επικεντρωθήκαμε στην αναζήτηση νομικών δεδομένων, που να πληρούν τις προδιαγραφές των Ανοιχτών Συνδεδεμένων Δεδομένων. Στην συνέχεια, διερευνήσαμε τα projects διαφόρων χωρών για την ελεύθερη πρόσβαση σε νομικά δεδομένα, ώστε να εκτελέσουμε τα ερωτήματα. Έχοντας ελέγξει τα www.legislation.gov.uk και το https://eur-lex.europa.eu διαπιστώσαμε ότι δεν υπάρχουν τα κατάλληλα SPARQL endpoints που να λειτουργούν και να επιστρέφουν αποτελέσματα, ώστε να εκτελέσουμε τα ερωτήματά μας. Παραδείγματος χάριν στο www.legislation.gov.uk, το οποίο είναι τεσσάρων αστέρων στο σχήμα του Tim Berners-Lee, είναι δυνατή μόνο η περιήγηση στα δεδομένα μέσω API και δεν παρέχεται η δυνατότητα να κάνουμε λήψη και μεταφόρτωση των δεδομένων ως σύνολο. Από την έρευνα μας καταλήξαμε ότι το Semantic Finlex [https://data.finlex.fi/en] έχει την κατάλληλη δομή για την εξόρυξη και ανάλυση των νομικών δεδομένων, διαθέτοντας SPARQL endpoint για την εκτέλεση ερωτημάτων καθώς και αναλυτική τεκμηρίωση του RDF μοντέλου για τη νομοθεσία και τη νομολογία.

Κατόπιν αυτού, μελετήσαμε την οντολογία του Semantic Finlex που δημιουργήθηκε για τη νομοθεσία και τη νομολογία (οντολογία SFL και SFCL αντίστοιχα). Ο επόμενος στόχος μας ήταν να διαγνώσουμε τη δομή ενός εγγράφου νομοθεσίας με την οντολογία SFL και αντίστοιχα τη δομή ενός εγγράφου νομολογίας με την οντολογία SFCL, προκειμένου να συντάξουμε το κατάλληλο query που θα επιστρέψει ενδιαφέρουσες πληροφορίες. Πιο συγκεκριμένα, μελετήσαμε τις κλάσεις και τις ιδιότητες τόσο των νόμων (Statute) όσο και των δικαστικών αποφάσεων (judgements) και καταλήξαμε στις μεταβλητές που θα μας βοηθήσουν στην ανάλυση των δεδομένων. Ακολούθως, συντάξαμε queries ζητώντας να επιστραφούν μόνο οι επιλεχθείσες μεταβλητές. Στην συνέχεια, χρησιμοποιήσαμε το online εργαλείο YASGUI [http://about.yasgui.org/] που είναι μια φιλική προς το χρήση διεπαφή για την εκτέλεση των SPARQL queries για

νομοθεσία και νομολογία, αντιγράφοντας το κείμενο του εκάστοτε ερωτήματος που συντάξαμε μέσα στο πλαίσιο ερωτήματος YASGUI, πληκτρολογώντας τη διεύθυνση http://data.finlex.fi/sparql στο πεδίο τελικού σημείου, σύμφωνα και με τις οδηγίες του Semantic Finlex. Κατόπιν αυτών, λάβαμε τα προς ανάλυση δεδομένα όπως στην συνέχεια παρατίθενται και προχωρήσαμε στην αξιολόγηση των αποτελεσμάτων.

Πέραν όμως των νομικών δεδομένων και επειδή στόχος της εργασίας ήταν να παρουσιάσουμε και τη διασύνδεση μεταξύ των ανοιχτών συνόλων δεδομένων με την χρήση των κατάλληλων τεχνικών εργαλείων για να τα αναλύσουμε και να τα εκμεταλλευτούμε προς εξαγωγή γνώσης, προχωρήσαμε στα ακόλουθα βήματα. Ανατρέξαμε σε στατιστικά σετ δεδομένων και καταλήξαμε στο statista (https://www.statista.com/), στο οποίο κάναμε δωρεάν εγγραφή προκειμένου να αποκτήσουμε πρόσβαση σε πρόσφατες έρευνες. Κάνοντας περιήγηση στα στατιστικά στοιχεία ειδικότερα χωρών και εταιριών εντοπίσαμε μια έρευνα που απεικονίζει στατιστικά στοιχεία χωρών σχετικά με τις αφίξεις τουριστών (σε εκατομμύρια) και το ποσό που αυτοί ξόδεψαν (σε δισεκατομμύρια δολάρια). Το επόμενο βήμα μας ήταν να εισάγουμε τα δεδομένα στο RapidMiner και να τα συνδέσουμε με δεδομένα από την DBpedia χρησιμοποιώντας τους κατάλληλους operators ώστε να ανακτήσουμε τα Uri's των χωρών αυτών. Τέλος, μέσω αυτών των Uris προστέθηκαν επιπλέον μεταβλητές έτσι ώστε να καταλήξουμε στην ανάλυση των αποτελεσμάτων όπως ακολουθεί στο τέταρτο κεφάλαιο της εργασίας.

3.5 Εκμάθηση της επέκτασης Ανοιχτών Συνδεδεμένων Δεδομένων [LOD extension] του RapidMiner

Η Επέκταση Ανοιχτών Συνδεδεμένων Δεδομένων στο RapidMiner [εφεξής RapidMiner LOD extension] δημιουργήθηκε από μία ομάδα ερευνητών του Πανεπιστήμιου του Mannheim υπό την καθοδήγηση του καθηγητή Heiko Paulheim. Με το LOD extension μπορούμε να αντλήσουμε δεδομένα από το LOD cloud και να τα εκμεταλλευτούμε για την ανάλυση και λήψη στρατηγικών αποφάσεων, την πρόβλεψη μελλοντικών καταστάσεων, την ερμηνεία των στατιστικών στοιχείων με την πρόσθεση νέων χαρακτηριστικών από τα datasets Ανοιχτών Συνδεδεμένων Δεδομένων και την ανάλυση των συσχετίσεων που προκύπτουν και την ταξινόμηση των δεδομένων σε κατηγορίες προς διευκόλυνση του χρήστη [*Janssen, Fallahi, Nobner & Paulheim, 2012*]. Επιπλέον,

η RapidMiner LOD extension μπορεί να λειτουργήσει με εντελώς αυθαίρετο τρόπο, πράγμα που σημαίνει ότι οι υποψήφιοι χρήστες δεν απαιτείται να διαθέτουν προχωρημένες σχετικά με τις πηγές των δεδομένων ή και τις τεχνολογίες όπως το RDF και το SPARQL για να το χρησιμοποιήσουν.

Για την εκμάθηση και την εκτέλεση των βημάτων χρησιμοποιήθηκε το εγχειρίδιο της έκδοσης 1.5 του RapidMiner LOD extension [*Paulheim, Ristoski, Mitichkin & Bizer, 2014]*, σύμφωνα με το οποίο η διαδικασία που ακολουθείται για την εξόρυξη και ανάλυση των δεδομένων περιλαμβάνει: (i) Εισαγωγή των Δεδομένων (Data Import), (ii) Σύνδεση των δεδομένων (Data Linking), (iii) Δημιουργία χαρακτηριστικών (Feature Generation), (iv) Επιλογή υποσυνόλου χαρακτηριστικών (Feature Subset Selection), (v) Εξερεύνηση των συνδέσμων (Exploring Links), (vi) Ενσωμάτωση των δεδομένων (Data Integration).

3.6 Συμπεράσματα

Από τα ανωτέρω προκύπτει ότι η μεθοδολογία μας έχει έξι βήματα. Μελέτη της βιβλιογραφίας, εκμάθηση της γλώσσας SPARQL για την εκτέλεση των ερωτημάτων, εξερεύνηση των ανοιχτών συνόλων νομικών δεδομένων, επιλογή του Semantic Finlex και μελέτη της οντολογίας του για σύνταξη queries και ανάλυση, επιλογή στατιστικού συνόλου δεδομένων προς ανάλυση και εκμάθηση της RapidMiner LOD extension. Την μεγαλύτερη σημασία στη μεθοδολογία μας είχε να ανακαλύψουμε σύνολα δεδομένων που είναι δομημένα σύμφωνα με τις αρχές των Ανοιχτών Συνδεδεμένων Δεδομένων και παρέχουν παράλληλα τη δυνατότητα εκτέλεσης ερωτημάτων με SPARQL σε endpoint που λειτουργούν. Η διαδικασία ανάκτησης και ανάλυσης των συνόλων δεδομένων που τελικώς επιλέξαμε βρίσκεται στο επίκεντρο της μελέτης μας.

4. ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ - ΕΡΜΗΝΕΙΑ ΤΩΝ ΕΥΡΗΜΑΤΩΝ

4.1 Εισαγωγή

Όπως αναφέραμε στο παραπάνω κεφάλαιο της μεθοδολογίας, επιλέξαμε να αντλήσουμε τα προς ανάλυση δεδομένα μας από το Semantic Finlex και από το statista. Επειδή θα χρησιμοποιήσουμε δύο διαφορετικές κατηγορίες δεδομένων για την εξαγωγή χρήσιμων συμπερασμάτων (νομικά και στατιστικά αντίστοιχα), είναι σημαντικό να επεξηγήσουμε τις τεχνικές και τις οντολογίες, να παρουσιάσουμε τις κλάσεις και τις ιδιότητες, να αιτιολογήσουμε την επιλογή συγκεκριμένων μεταβλητών και να περιγράψουμε τη διαδικασία εξόρυξης και ανάλυσης. Προς διευκόλυνση του αναγνώστη, το παρόν κεφάλαιο χωρίζεται σε δύο ενότητες: η πρώτη περιλαμβάνει τα δύο παραδείγματα εξόρυξης και ανάλυσης νομικών δεδομένων από το Semantic Finlex και την ερμηνεία των ευρημάτων και η δεύτερη παρουσιάζει τα πορίσματα της εξόρυξης και διασύνδεσης δεδομένων από το statista με το RapidMiner LOD extension και την ανάλυση αυτών.

4.2 Η διαδικασία απόκτησης δεδομένων από το Semantic FinLex

Το Semantic Finlex περιέχει τα σύνολα δεδομένων της ενοποιημένης νομοθεσίας και των πρωτότυπων κειμένων, πράξεων και διαταγμάτων της Φιλανδίας σε μορφή RDF. Τα πρωτότυπα κείμενα (sfl:Original) περιέχουν την αρχική μορφή των νόμων κατά την ψήφιση και δημοσίευσή τους, ενώ η ενοποιημένη νομοθεσία (sfl:Consolidated) περιέχει ενημερωμένες εκδόσεις των αρχικών κειμένων. Το κείμενο του κάθε νόμου επικαροποιείται κάθε φορά που δημοσιεύεται μια τροποποίηση. Η ομάδα υλοποίησης του Semantic Finlex δημιούργησε καταρχήν το σχήμα Finlex XML και ως βασικές κατηγορίες RDFS του μοντέλου τα Statute (για τη νομοθεσία) και CourtDecision (για τις δικαστικές αποφάσεις). Περαιτέρω, η ομάδα του έργου αξιολόγησε εάν το πρότυπο ELI, το οποίο εφαρμόζει οντολογικά το μοντέλο FRBR των βιβλιοθηκών, παρέχει τη δυνατότητα διάγνωσης και επίλυσης των δυσχερειών που προκύπτουν από τα ειδικά χαρακτηριστικά των νομικών κειμένων. Το συγκεκριμένο μοντέλο εκτεταμένων δεδομένων είναι εμπνευσμένο από την ορολογία του FRBR, όπου το «έργο» αναφέρεται στο έγγραφο σε εννοιολογικό επίπεδο (ως χωριστή γραπτή δημιουργία), η «έκφραση» αναφέρεται σε ξεχωριστή υλοποίηση ενός έργου (για παράδειγμα, σε μια συγκεκριμένη γλωσσική έκδοση) και η «εκδήλωση» αναφέρεται στη φυσική εκδήλωση μιας έκφρασης

(μια συγκεκριμένη δημοσίευση, όπως ένα PDF). Η ELI εφαρμόζει το μοντέλο FRBR δημιουργώντας δικά του αντίγραφα για νομικά σύνολα δεδομένων ως εξής:

→eli:LegalResource που αντιστοιχεί στο επίπεδο του έργου, πχ στο νόμο σε εννοιολογικό επίπεδο
→eli:LegalExpression που αντιστοιχεί στο επίπεδο έκφρασης, δηλαδή στη γλωσσική έκδοση ενός νόμου
→eli:format που αντιστοιχεί στο επίπεδο εκδήλωσης, δηλαδή στη γλωσσική έκδοση σε μια συγκεκριμένη μορφή περιεχομένου

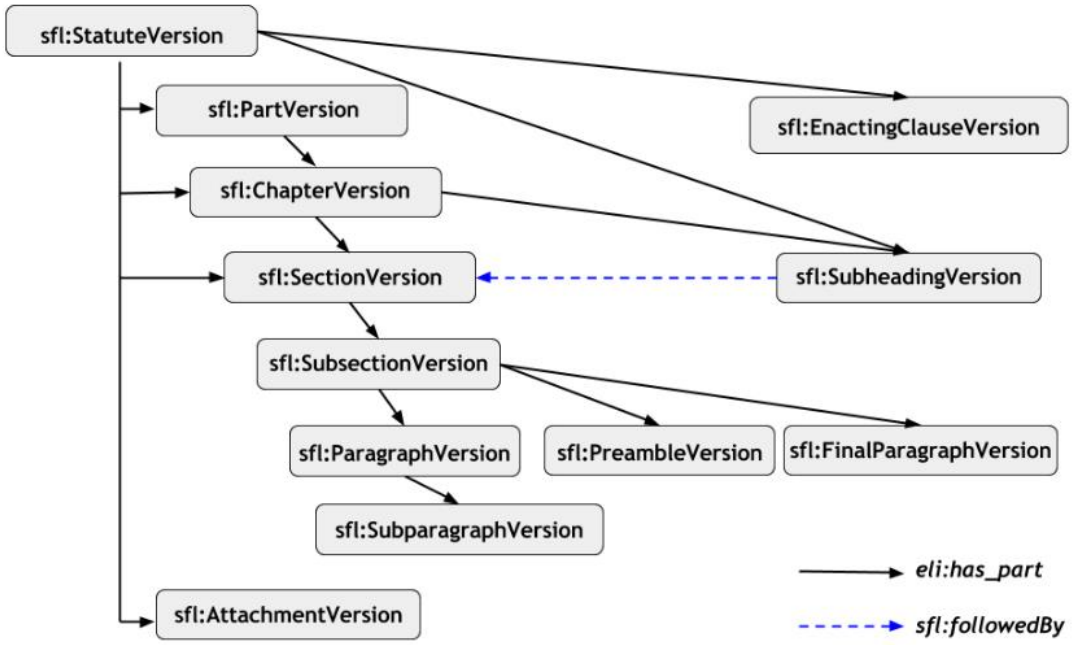
4.2.1 Περίπτωση πρώτη: ερώτημα για άντληση δεδομένων νομοθετημάτων

Η ομάδα υλοποίησης του Semantic Finlex **κατέληξε στην επέκταση της οντολογίας ELI για τη νομοθεσία με την οντολογία SFL** (Semantic Finlex Legislation) που ορίζει ξεχωριστές κλάσεις για την κάθε οντότητα. Εκτός από τις διαφορετικές εκδόσεις μορφής γλώσσας και περιεχομένου, η SFL ορίζει κατηγορίες για διαφορετικές χρονικές εκδόσεις των νομοθετημάτων. Επομένως, η SFL **διαιρεί την κλάση eli: LegalResource** σε τέσσερις υποκατηγορίες ως εξής:

→sfl: Statute. Νομοθέτημα σε εννοιολογικό επίπεδο π.χ. Ποινικός κώδικας (39/1889)
→sfl: SectionOfALaw. Τμήμα νομοθετήματος σε εννοιολογικό επίπεδο, π.χ. Ποινικός κώδικας 39/1889 Κεφάλαιο 1 Τμήμα 2
→sfl: StatuteVersion. Νομοθέτημα σε συγκεκριμένο χρονικό σημείο, π.χ. Ποινικός Κώδικας 39/1889 όπως ήταν στις 1.1.2016
→sfl: SectionOfALawVersion. Τμήμα νομοθετήματος σε συγκεκριμένο χρονικό σημείο, π.χ. Ποινικός Κώδικα 39/1889 Κεφάλαιο 1 Τμήμα 2 όπως ισχύει την 1.1.2016

Η σύνδεση των νόμων και των επιμέρους τμημάτων τους με τις χρονικές εκδόσεις τους πραγματοποιείται με τις ιδιότητες eli:has_member και eli:is_member_of. Η σύνδεση μεταξύ ενός τμήματος νομοθετήματος σε εννοιολογικό επίπεδο και ενός τμήματος νομοθετήματος σε συγκεκριμένο χρονικό σημείο εκφράζεται παρομοίως με τις ίδιες ιδιότητες.

Εικόνα 17: Η δομή ενός εγγράφου νομοθεσίας με την οντολογία SFL



Η δομή ενός εγγράφου [εικόνα 17] διαφέρει από τις χρονικές εκδόσεις του. Για παράδειγμα, μεμονωμένα μέρη ενδέχεται να προστεθούν, να αφαιρεθούν, να αλλαχθούν ή να μεταφερθούν.

Ακολούθως μελετήσαμε προσεκτικά τις κλάσεις και τις ιδιότητες [εικόνες 18-23] προκειμένου να προχωρήσουμε στην επιλογή των κατάλληλων μεταβλητών και να συντάξουμε το ερώτημα. Οι κλάσεις και οι ιδιότητες ανακτήθηκαν από το https://data.finlex.fi/en/legislation και τα χαρακτηριστικά τους (αναγνωριστικό, εύρος τιμών, συσχέτιση, τύπος τιμών και περιγραφή) παρουσιάζονται στις ακόλουθες εικόνες:

Εικόνα 18: Οι κλάσεις και οι ιδιότητες του sfl:Statute

sfl:Statute (statute)

Identifier	Range	Cardinality	Value type	Description
eli:has_member	*sfl:StatuteVersion*	[0, n]	URI	statute version
eli:id_local	*rdfs:Literal*	1	Literal	statute id #/YYYY
eli:type_document	• *sfl:Act* • *sfl:PresidentialDecree* • *sfl:GovernmentDecree* • *sfl:MinisterialDecree* • *sfl:GovernmentDecision* • *sfl:OtherDecision* • *sfl:OtherStatute*	1	URI	legislative level
sfl:statuteType	• *sfl:NewStatute* • *sfl:Amendment* • *sfl:Repeal*	1	URI	statute type
eli:in_force	• *eli:InForce-notInForce*	[0,1]	URI	is not in force

Εικόνα 19: Οι κλάσεις και οι ιδιότητες του sfl:SectionOfALaw

sfl:SectionOfALaw (section of a law)

Subclasses: *sfl:Part, sfl:Chapter, sfl:Section, sfl:Subsection, sfl:Paragraph, sfl:Subparagraph, sfl:EnactingClause, sfl:Attachment, sfl:Subheading, sfl:Preamble, sfl:FinalParagraph, sfl:EntryIntoForce*

Identifier	Range	Cardinality	Value type	Description
eli:has_member	sfl:SectionOfALawVersion	[0, n]	URI	temporal version of a section of a law
eli:in_force	• *eli:InForce-notInForce*	[0,1]	URI	is not in force

Εικόνα 20: Οι κλάσεις και οι ιδιότητες του sfl:StatuteVersion

sfl:StatuteVersion

Identifier	Range	Cardinality	Value type	Description
eli:is_member_of	*sfl:Statute*	[0,n]	URI	version of a statute
eli:has_part	*sfl:SectionOfALawVersion*	[0,n]	URI	contains section of a law
eli:is_realized_by	*eli:LegalExpression*	2	URI	language version
eli:passed_by	*eli:Agent*	1	URI	person/organization that originally passed the law
eli:is_about	*skos:Concept*	[0,n]	URI	a subject for the statute
eli:first_date_entry_in_force	*xsd:date*	1	Literal	date of entry into force
eli:date_document	*xsd:date*	1	Literal	date of signature
eli:date_publication	*xsd:date*	1	Literal	date of publication
eli:version	• *sfl:Consolidated* (consolidated, i.e. amended) • *sfl:Original* (original)	1	URI	version
eli:version_date	*xsd:date*	1	URI	version date (date of entry into force of the amendment)
eli:amended_by	*sfl:StatuteVersion*	[0,1]	URI	amended by the statute
eli:responsibility_of	*rdfs:Literal*	1	Literal	person/organization responsible for the statute
eli:related_to	*sfl:GovernmentProposal*	[0,n]	URI	government proposal
eli:based_on	*sfl:Statute*	[0,n]	URI	based on a statute
eli:amends	*sfl:Statute*	[0,n]	URI	amends statute
eli:repeals	*sfl:Statute*	[0,n]	URI	repeals statute
eli:repealed_by	*sfl:Statute*	[0,1]	URI	repealed by statute

- Original versions
- Consolidated (amended) versions

Εικόνα 21: Οι κλάσεις και οι ιδιότητες του sfl:SectionOfALawVersion

sfl:SectionOfALawVersion

Subclasses: *sfl:PartVersion, sfl:ChapterVersion, sfl:SectionVersion, sfl:SubsectionVersion, sfl:ParagraphVersion, sfl:SubparagraphVersion, sfl:EnactingClauseVersion, sfl:AttachmentVersion, sfl:SubheadingVersion, sfl:PreambleVersion*

Identifier	Range	Cardinality	Value type	Description
eli:is_member_of	*sfl:SectionOfALaw*	1	URI	is version of a section of a law
eli:has_part	*sfl:SectionOfALawVersion*	[0,n]	URI	contains a section of a law
eli:is_part_of	• *sfl:StatuteVersion* (statute version) • *sfl:SectionOfALawVersion* (section of a law version)	[0,n]	URI	is part of a statute / section of a law
eli:version	• *sfl:Consolidated* (consolidated, i.e. amended) • *sfl:Original* (original)	1	URI	version
eli:version_date	*xsd:date*	1	URI	date of entry into force of the amendment
eli:amended_by	*sfl:StatuteVersion*	[0,1]	URI	changed by a statute
eli:cites	*sfl:Statute*	[0,n]	URI	in text reference to a statute / section of a law
sfl:followedBy **	*sfl:SectionVersion*	1	URI	section following a subheading

- *sfl:Consolidated* (consolidated, i.e. amended)
- *sfl:Original* (original)
- ** Subheading version (*sfl:SubheadingVersion*)

Εικόνα 22: Οι κλάσεις και οι ιδιότητες του eli:LegalExpression για την επιλογή γλώσσας νομοθετήματος (φιλανδικά ή σουηδικά ή και τα δύο)

eli:LegalExpression (language version)

Identifier	Range	Cardinality	Value type	Description
eli:realizes	• *sfl:StatuteVersion* • *sfl:SectionOfALawVersion*	1	URI	statute/section of a law temporal version
eli:is_embodied_by	*eli:Format*	[2,3]	URI	content format
eli:language	*eli:Language*	1	Literal	language
eli:title	*rdfs:Literal*	1	Literal	title
eli:title_alternative	*rdfs:Literal*	[0,n]	Literal	alternative title

Εικόνα 23: Οι κλάσεις και οι ιδιότητες του eli:Format (οι διαφορετικές μορφές περιεχομένου πχ κείμενο και HTML των παραλλαγών γλώσσας μοντελοποιούνται ως στιγμιότυπα της κλάσης eli: Format και συνδέονται με την ιδιότητα eli:embodies).

eli:Format (content format)

Identifier	Range	Cardinality	Value type	Description
eli:embodies	*eli:LegalExpression*	1	URI	language version
eli:format		1	URI	content format (see here)
sfl:text	*rdfs:Literal*	1	Literal	content in text format
sfl:html	*rdfs:Literal*	1	HTML	content in HTML format

Έχοντας κατανοήσει την οντολογία και τη δομή των νομοθετημάτων του Semantic Finlex, καταλήξαμε στη σύνταξη ενός ερωτήματος που θα μας επιστρέψει τα πέντε νομοθετήματα (νόμους, προεδρικά διατάγματα, υπουργικές αποφάσεις κλπ) που στο σύνολό τους (και όχι ένα συγκεκριμένο τμήμα τους) έχουν τροποποιηθεί τις περισσότερες φορές (έχουν το μεγαλύτερο πλήθος versions) και είναι σε ισχύ.

Μέσω του εργαλείου http://yasgui.org/# ορίσαμε ως endpoint το http://data.finlex.fi/sparql. Στην συνέχεια συντάξαμε το ακόλουθο ερώτημα:

```
PREFIX xs: <http://www.w3.org/2001/XMLSchema#>

PREFIX http: <http://www.w3.org/2011/http#>

PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX sfcl: <http://data.finlex.fi/schema/sfcl/>

PREFIX sfl: <http://data.finlex.fi/schema/sfl/>

PREFIX eli: <http://data.europa.eu/eli/ontology#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

PREFIX http: <http://www.w3.org/2011/http#>


# Query : List of Statutes

SELECT ?statute ?type_document ?date_document (max(?statute_versiondate) as

?LastVersionDate) (min(?statute_versiondate) as ?FirstVersionDate)
(count(?statuteVersion) as ?NumberOfVersions)

WHERE {

?statute rdf:type sfl:Statute ;

eli:type_document ?type_document;

eli:has_member ?statuteVersion.
```

```
FILTER NOT EXISTS { ?statute eli:in_force eli:InForce-notInForce . }

?statuteVersion eli:version_date ?statute_versiondate;

eli:date_document ?date_document.

FILTER (?statute = <http://data.finlex.fi/eli/sd/2015/379> )

}

group by ?statute ?type_document ?date_document

order by desc(?NumberOfVersions)

LIMIT 5
```

Το ερώτημα αναζητά τα νομοθετήματα (?statute), το είδος του εγγράφου (?type_document), την ημερομηνία υπογραφής (?date_document) ζητώντας την ημερολογιακά πιο πρόσφατη τροποποίηση (?MaxVersionDate), την ημερολογιακά πιο παλιά τροποποίηση (?MinVersionDate) καθώς και το πλήθος των τροποποιήσεων του κάθε νομοθετήματος (?CountOfVersions). Στο συγκεκριμένο ερώτημα περιορίζουμε τα αποτελέσματά μας μέσω φίλτρων ορίζοντας να είναι σε ισχύ το νομοθέτημα μέσω της ιδιότητας eli:in_force και τέλος, προκειμένου να κάνουμε έλεγχο ορθής απεικόνισης των αποτελεσμάτων περιορίζουμε τη αναζήτηση σε ένα συγκεκριμένο νομοθέτημα (http://data.finlex.fi/eli/sd/2015/379). Απαραίτητη προϋπόθεση για την επιτυχή εκτέλεση του ερωτήματος ήταν ο ορισμός προθεμάτων (prefixes) που περιέχουν τους χώρους ονομάτων (namespaces) που χρησιμοποιήθηκαν πχ eli, sfl, rdf, skos κλπ.

Αμέσως παρακάτω είναι το ερώτημα στο γραφικό περιβάλλον του yasgui.

http://data.finlex.fi/sparql

```
PREFIX xs: <http://www.w3.org/2001/XMLSchema#>
PREFIX http: <http://www.w3.org/2011/http#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sfcl: <http://data.finlex.fi/schema/sfcl/>
PREFIX sfl: <http://data.finlex.fi/schema/sfl/>
PREFIX eli: <http://data.europa.eu/eli/ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX http: <http://www.w3.org/2011/http#>

# Query : List of Statutes
SELECT ?statute ?type_document ?date_document (max(?statute_versiondate) as
?LastVersionDate) (min(?statute_versiondate) as ?FirstVersionDate) (count(?statuteVersion) as ?NumberOfVersions)
WHERE {
?statute rdf:type sfl:Statute ;
eli:type_document ?type_document;
eli:has_member ?statuteVersion.
FILTER NOT EXISTS { ?statute eli:in_force eli:InForce-notInForce . }
?statuteVersion eli:version_date ?statute_versiondate;
eli:date_document ?date_document.
FILTER (?statute = <http://data.finlex.fi/eli/sd/2015/379> )
}
group by ?statute ?type_document ?date_document
order by desc(?NumberOfVersions)
LIMIT 5
```

Εκτελούμε το ερώτημα και το αποτέλεσμα [εικόνα 24] είναι το εξής:

Εικόνα 24: Αποτέλεσμα query

Table | Response | Pivot Table | Google Chart | Geo

Showing 1 to 1 of 1 entries (in 0.515 seconds)

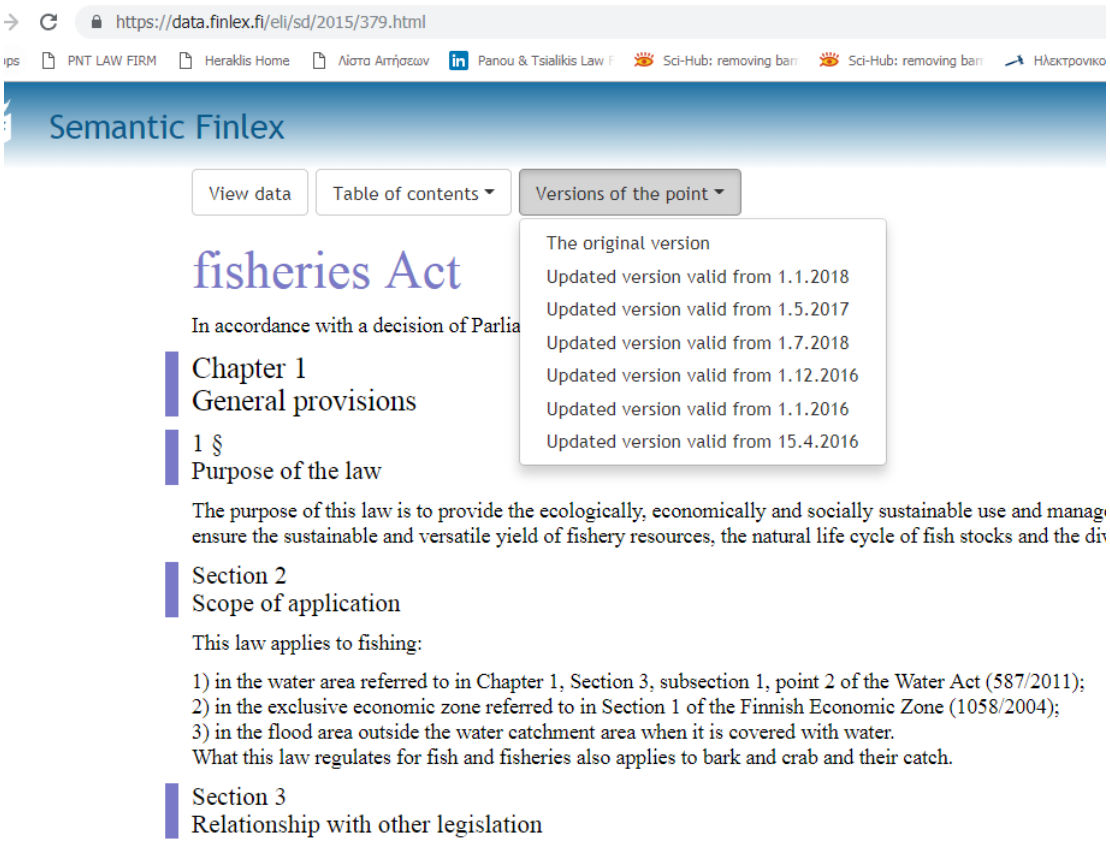
Search: Show 50

	statute	type_document	date_document	LastVersionDate	FirstVersionDate	NumberOfVersions	Dif_years	Frequence_Cons
1	http://data.finlex.fi/eli/sd/2015/379	http://data.finlex.fi/schema/sfl/Act	"2015-04-10"^^xsd:date	"2018-07-01"^^xsd:date	"2016-01-01"^^xsd:date	"6"^^xsd:integer	"3"^^xsd:integer	"0.5"^^xsd:decimal

Showing 1 to 1 of 1 entries (in 0.515 seconds)

Προκειμένου να ελέγξουμε την ακρίβεια του ερωτήματος, δηλαδή να επιβεβαιώσουμε ότι εμφανίστηκαν τα σωστά αποτελέσματα αναφορικά με το πλήθος των εκδόσεων ενός statute, φιλτράραμε τα αποτελέσματα σε ένα συγκεκριμένο statute (νομοθέτημα 2015/379) όπως αναφέραμε και παραπάνω. Κάνοντας κλικ στο URI στη στήλη statute, εμφανίζεται το νομοθέτημα σε νέα σελίδα σε HTML μορφή [εικόνα 25].

Εικόνα 25: Πρώτος Έλεγχος (έλεγχος πλήθους εκδόσεων σε html format)



Έχοντας κάνει αυτόματη μετάφραση της σελίδας που εμφανίζεται από τα Φιλανδικά στα Αγγλικά, παρατηρούμε 6 διαφορετικές εκδόσεις του παραπάνω νομοθετήματος (2015/379) εκτός της αρχικής (original) με ημερομηνίες έκδοσης τις εξής:

01.01.2018

01.05.2017

01.07.2018

01.12.2016

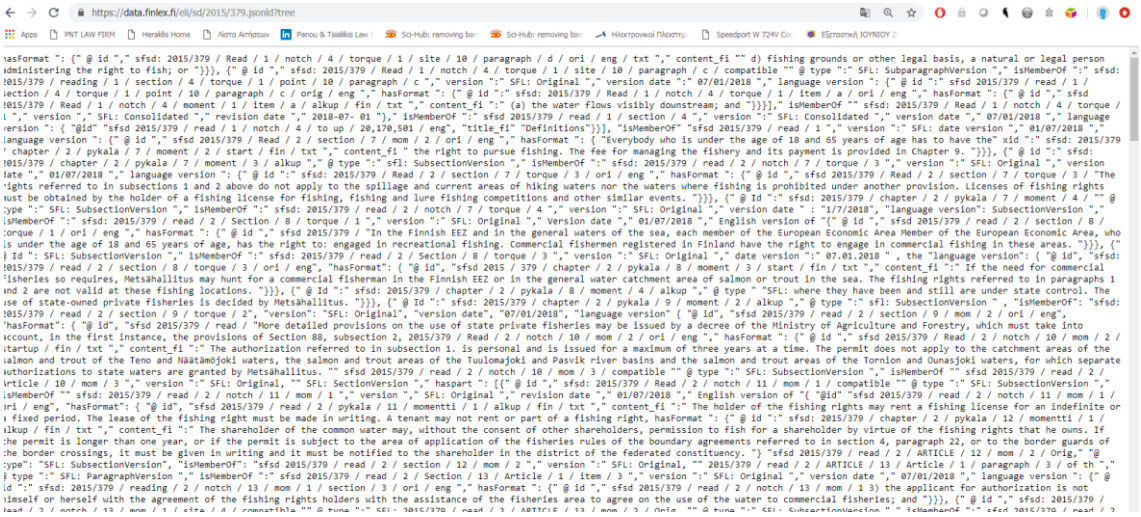
01.01.2016

15.04.2016

Επιβεβαιώνεται επίσης ότι η πιο πρόσφατη έκδοση έχει δημοσιευθεί την 01.07.2018 (MaxVersionDate) καθώς και ότι η $1^{η}$ τροποποίηση (MinVersionDate) πραγματοποιήθηκε την 01.01.2016.

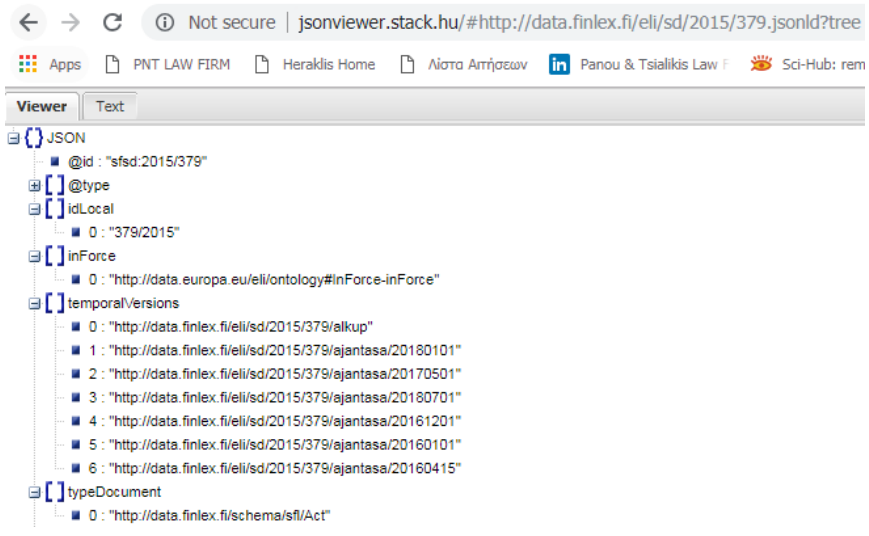
Παρατηρώντας την Html μορφή του Εγγράφου, μας δίνεται η επιλογή να ανακτήσουμε τα δεδομένα του σε μορφή JSON [εικόνα 26] κάνοντας κλικ στην επιλογή View Data [https://data.finlex.fi/eli/sd/2015/379.jsonld?tree].

Εικόνα 26: Δεύτερος έλεγχος (μέσω JSON tree). Δεδομένα νομοθετήματος 2015/379 σε μορφή JSON.



Συμπεραίνουμε ότι δεν μας βοηθάει ιδιαίτερα η παραπάνω JSON μορφή του εγγράφου και για τον λόγο αυτό θα κάνουμε χρήση ενός εργαλείου JSON viewer (http://jsonviewer.stack.hu/), εισάγοντας το Url του JSON δέντρου το οποίο θέλουμε να απεικονίσουμε και κάνοντας επέκταση των ιδιοτήτων της κλάσης προκειμένου να εντοπίσουμε και πάλι το πλήθος των εκδόσεων του συγκεκριμένου νομοθετήματος [εικόνα 27].

Εικόνα 27: Εισαγωγή δεδομένων νομοθετήματος 2015/379 σε JSON viewer



Επιβεβαιώνεται για 2η φορά η ορθότητα του ερωτήματος που συντάξαμε και πλέον μπορούμε να το εκτελέσουμε στο σύνολό του [εικόνα 28], απαλείφοντας δηλαδή το φίλτρο του νομοθετήματος 2015/379, έτσι ώστε να αντλήσουμε τα 5 πρώτα σε τροποποιήσεις νομοθετήματα που είναι σε ισχύ, κατά φθίνουσα σειρά ως προς το πλήθος των τροποποιήσεων (order by desc(?CountOfVersions) & Limit 5), απεικονίζοντας τα εξής στοιχεία:

- Uri νομοθετήματος
- Τύπος Εγγράφου (Act, Presidential Decree, Government Decree, Ministerial Decree, Government Decision, Other Statute, Other Decision).
- Ημερομηνία Εγγράφου
- Πιο πρόσφατη ημερομηνία τροποποίησης.
- Πρώτη ημερομηνία τροποποίησης.
- Πλήθος διαφορετικών εκδόσεων.

Εικόνα 28: Το ερώτημα για τα πέντε νομοθετήματα με τις περισσότερες τροποποιήσεις

http://data.finlex.fi/sparql

```
PREFIX http: <http://www.w3.org/2011/http#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sfcl: <http://data.finlex.fi/schema/sfcl/>
PREFIX sfl: <http://data.finlex.fi/schema/sfl/>
PREFIX eli: <http://data.europa.eu/eli/ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX http: <http://www.w3.org/2011/http#>

# Query : List of Statutes
SELECT ?statute ?type_document ?date_document (max(?statute_versiondate) as
?LastVersionDate) (min(?statute_versiondate) as ?FirstVersionDate) (count(?statuteVersion) as ?NumberOfVersions)

WHERE {
?statute rdf:type sfl:Statute ;
eli:type_document ?type_document;
eli:has_member ?statuteVersion.
FILTER NOT EXISTS { ?statute eli:in_force eli:InForce-notInForce . }
?statuteVersion eli:version_date ?statute_versiondate;
eli:date_document ?date_document.
}
group by ?statute ?type_document ?date_document ?dif_years ?version_rate
order by desc(?NumberOfVersions)
LIMIT 5
```

Εικόνα 29: Αποτέλεσμα ερωτήματος με φθίνουσα σειρά ως προς το πλήθος των εκδόσεων

Showing 1 to 5 of 5 entries

	statute	type_document	date_document	LastVersionDate	FirstVersionDate	NumberOfVersions
1	http://data.finlex.fi/eli/sd/1889/39	http://data.finlex.fi/schema/sfl/Act	"1889-12-19"^^xsd:date	"2018-08-15"^^xsd:date	"2016-01-01"^^xsd:date	"17"^^xsd:integer
2	http://data.finlex.fi/eli/sd/2008/878	http://data.finlex.fi/schema/sfl/Act	"2008-12-19"^^xsd:date	"2018-06-05"^^xsd:date	"2015-11-26"^^xsd:date	"15"^^xsd:integer
3	http://data.finlex.fi/eli/sd/2004/301	http://data.finlex.fi/schema/sfl/Act	"2004-04-30"^^xsd:date	"2018-04-01"^^xsd:date	"2015-08-01"^^xsd:date	"12"^^xsd:integer
4	http://data.finlex.fi/eli/sd/2014/527	http://data.finlex.fi/schema/sfl/Act	"2014-06-27"^^xsd:date	"2018-08-15"^^xsd:date	"2016-01-01"^^xsd:date	"11"^^xsd:integer
5	http://data.finlex.fi/eli/sd/1995/1558	http://data.finlex.fi/schema/sfl/Act	"1995-12-18"^^xsd:date	"2018-05-01"^^xsd:date	"2016-01-01"^^xsd:date	"11"^^xsd:integer

Showing 1 to 5 of 5 entries

Τα πέντε νομοθετήματα που προέκυψαν[εικόνα 29] είναι:

http://data.finlex.fi/eli/sd/1889/39 1889/39 -> Ποινικός Νόμος

http://data.finlex.fi/eli/sd/2008/878 2008/878 -> Νόμος περί Οικονομικής Επιτήρησης

http://data.finlex.fi/eli/sd/2004/301 2004/301 -> Νόμος περί αλλοδαπών

http://data.finlex.fi/eli/sd/1995/1558 1995/1558 -> Φορολογικός νόμος

http://data.finlex.fi/eli/sd/2014/527 2014/527 -> Νόμος για την προστασία του περιβάλλοντος

Από το αποτέλεσμα του ερωτήματος, συμπεραίνουμε καταρχήν ότι το νομοθέτημα με τις περισσότερες τροποποιήσεις (17), δηλαδή ο Ποινικός Νόμος, είναι και το παλαιότερο, καθώς η αρχική του έκδοση ανατρέχει στις 19.12.1889. Ο μεγάλος αριθμός των τροποποιήσεων είναι δικαιολογημένος για λόγους επικαιροποίησης και εναρμόνισης με τις παρούσες συνθήκες αλλά και για να συμπεριληφθούν νέες μορφές αξιόποινων πράξεων όπως πχ εγκλήματα που διαπράττονται μέσω διαδικτύου. Επίσης, από τον μεγάλο αριθμό τροποποιήσεων των νόμων που σχετίζονται με την οικονομική επιτήρηση και τη φορολογία συμπεραίνεται ότι και στη Φιλανδία το οικονομικό περιβάλλον μεταβάλλεται διαρκώς. Απόλυτα δικαιολογημένες είναι και οι πολυάριθμες αλλαγές το νόμο περί αλλοδαπών λόγω της κοινώς γνωστής μεταναστευτικής κρίσης στη Ευρώπη αλλά και στο νόμο που αφορά στην προστασία του περιβάλλοντος εξαιτίας των ραγδαίων κλιματικών αλλαγών. Είναι γεγονός ότι θα μπορούσαμε να εξάγουμε περισσότερα και πιο ενδιαφέροντα συμπεράσματα, εάν υπήρχαν περισσότερες χώρες που εφάρμοζαν το μοντέλο του Semantic Finlex στα νομικά τους δεδομένα, καθώς έτσι θα ήταν δυνατή μια συγκριτική ανάλυση ως προς το πλήθος και την συχνότητα των τροποποιήσεων νόμων που αφορούν πχ τη φορολογία, τους αλλοδαπούς κλπ σε περισσότερες χώρες και μπορούσαμε περαιτέρω να διασυνδέσουμε τα αποτελέσματα με άλλα δεδομένα των

χωρών αυτών όπως πχ το ετήσιο ΑΕΠ για να εξετάσουμε κατά πόσο οι συχνές αλλαγές στη νομοθεσία επηρεάζουν τις καταναλωτικές συνήθειες, τις επενδύσεις κ.ο.κ.

Για να διαπιστώσουμε την συχνότητα με την οποία τροποποιούνται τα νομοθετήματα, προσθέσαμε στην συνέχεια δύο ακόμη μεταβλητές στο παραπάνω ερώτημα, ζητώντας να μας επιστραφούν τα έτη μεταξύ της ημερομηνίας υπογραφής του νόμου μέχρι και σήμερα «2018» (?dif_years) και ο ρυθμός τροποποιήσεων (?version_rate) [εικόνα 30].

Εικόνα 30: Το ερώτημα για τα πέντε νομοθετήματα με τη μεγαλύτερη συχνότητα τροποποιήσεων

http://data.finlex.fi/sparql

```
PREFIX xs: <http://www.w3.org/2001/XMLSchema#>
PREFIX http: <http://www.w3.org/2011/http#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sfcl: <http://data.finlex.fi/schema/sfcl/>
PREFIX sfl: <http://data.finlex.fi/schema/sfl/>
PREFIX eli: <http://data.europa.eu/eli/ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX http: <http://www.w3.org/2011/http#>

# Query : List of Statutes
SELECT ?statute ?type_document ?date_document (max(?statute_versiondate) as
?LastVersionDate) (min(?statute_versiondate) as ?FirstVersionDate) (count(?statuteVersion) as ?NumberOfVersions) ?dif_years
 ((?NumberOfVersions/?dif_years) as ?version_rate)

WHERE {
?statute rdf:type sfl:Statute ;
eli:type_document ?type_document;
eli:has_member ?statuteVersion.
FILTER NOT EXISTS { ?statute eli:in_force eli:InForce-notInForce . }
?statuteVersion eli:version_date ?statute_versiondate;
eli:date_document ?date_document.

BIND("2018"^^xsd:integer - year(?date_document) as ?dif_years)
}
group by ?statute ?type_document ?date_document ?dif_years ?version_rate
order by desc(?version_rate)
LIMIT 5
```

Εικόνα 31: Αποτέλεσμα ερωτήματος με φθίνουσα σειρά ως προς το ρυθμό τροποποίησης

Showing 1 to 5 of 5 entries (in 1.24 seconds)

Search:

	statute	type_document	date_document	LastVersionDate	FirstVersionDate	NumberOfVersions	dif_years	version_rate
1	http://data.finlex.fi/eli/sd/2017/320	http://data.finlex.fi/schema/sfl/Act	"2017-05-24"^^xsd:date	"2018-08-15"^^xsd:date	"2018-01-01"^^xsd:date	"5"^^xsd:integer	"1"^^xsd:integer	"5.0"^^xsd:decimal
2	http://data.finlex.fi/eli/sd/2015/234	http://data.finlex.fi/schema/sfl/GovernmentDecree	"2015-03-19"^^xsd:date	"2018-06-13"^^xsd:date	"2015-12-14"^^xsd:date	"9"^^xsd:integer	"3"^^xsd:integer	"3.0"^^xsd:decimal
3	http://data.finlex.fi/eli/sd/2017/445	http://data.finlex.fi/schema/sfl/Act	"2017-06-28"^^xsd:date	"2018-08-15"^^xsd:date	"2017-06-28"^^xsd:date	"3"^^xsd:integer	"1"^^xsd:integer	"3.0"^^xsd:decimal
4	http://data.finlex.fi/eli/sd/2014/527	http://data.finlex.fi/schema/sfl/Act	"2014-06-27"^^xsd:date	"2018-08-15"^^xsd:date	"2016-01-01"^^xsd:date	"11"^^xsd:integer	"4"^^xsd:integer	"2.75"^^xsd:decimal
5	http://data.finlex.fi/eli/sd/2014/610	http://data.finlex.fi/schema/sfl/Act	"2014-08-08"^^xsd:date	"2018-08-15"^^xsd:date	"2016-01-01"^^xsd:date	"10"^^xsd:integer	"4"^^xsd:integer	"2.5"^^xsd:decimal

Showing 1 to 5 of 5 entries (in 1.24 seconds)

Τα πέντε νομοθετήματα που προέκυψαν[εικόνα 31] είναι:

http://data.finlex.fi/eli/sd/2017/320 -> Νόμος για τις υπηρεσίες μεταφορών

http://data.finlex.fi/eli/sd/2015/234 -> Κυβερνητικό Διάταγμα για την ίδρυση, οικολογική στήριξη και υποστήριξη νέων αγροτών

http://data.finlex.fi/eli/sd/2017/445 -> Νόμος για το ξέπλυμα χρήματος από παράνομες δραστηριότητες

http://data.finlex.fi/eli/sd/2014/527 -> Νόμος για την προστασία του περιβάλλοντος

http://data.finlex.fi/eli/sd/2014/610 -> Νόμος για πιστωτικά ιδρύματα

Προκύπτει λοιπόν ότι το νομοθέτημα με το μεγαλύτερο ρυθμό τροποποιήσεων (πέντε φορές μέσα ένα χρόνο) ρυθμίζει τις μεταφορικές υπηρεσίες. Παρατηρούμε ότι εμφανίζεται πάλι ο νόμος που αφορά στην προστασία του περιβάλλοντος με ρυθμό τροποποίησης 2,75 (11 αλλαγές σε 4 έτη) και διαπιστώνουμε επίσης ότι υπάρχει συχνότητα τροποποιήσεων σε νόμους που αφορούν στην οικονομία (ξέπλυμα χρήματος, πιστωτικά ιδρύματα).

4.2.2 Περίπτωση δεύτερη: ερώτημα για άντληση δεδομένων δικαστικών αποφάσεων

Όπως προαναφέρθηκε, **και για την νομολογία αναπτύχθηκε η εξειδικευμένη οντολογία SFCL** (Semantic Finlex Case Law), ένα προσαρμοσμένο μοντέλο μεταδεδομένων που υλοποιήθηκε ως μέρος του έργου Semantic Finlex. Η SFCL εφαρμόζει το πρότυπο FRBR με παρόμοιο τρόπο όπως το ELI και το SFL στην περίπτωση της νομοθεσίας. Το μοντέλο μεταδεδομένων περιέχει τις ιδιότητες του Dublin Core που συμπεριλαμβάνονται στο πρότυπο ECLI. Ωστόσο, υπάρχει μόνο μία χρονική εκδοχή κάθε απόφασης και επομένως το μοντέλο FRBR της οντολογίας SFCL αποτελείται μόνο από τις ακόλουθες τρεις κλάσεις.

sfcl: Judgement (έργο): η δικαστική απόφαση αυτή καθαυτή

sfcl: Expression (έκδοση γλώσσας): γλωσσική έκδοση απόφασης

sfcl: Format (μορφή περιεχομένου): απόφαση σε συγκεκριμένη μορφή περιεχομένου

Εικόνα 32: Οι κλάσεις και οι ιδιότητες των δικαστικών αποφάσεων

sfcl:Judgment (judgment)

Identifier	Range	Cardinality	Value type	Description
dcterms:coverage	*dcterms:Location*	1	URI	the country of the court
dcterms:contributor	*dcterms:Agent*	[0,n]	Literal	judges
dcterms:creator	*dcterms:Agent*	1	URI	court
dcterms:date	*xsd:date*	1	Literal	date of judgment
dcterms:description	*skos:Concept*	[0,n]	URI	keywords
dcterms:issued	*xsd:date*	1	Literal	date of publication
dcterms:isVersionOf	*rdfs:Literal*	1	Literal	ECLI identifier of the judgment
dcterms:type	*skos:Concept*	1	URI	type of the decision
sfcl:isRealizedBy	*sfcl:Expression*	[0,n]	URI	language version

Εικόνα 33: Οι κλάσεις και οι ιδιότητες των γλωσσικών εκδόσεων των αποφάσεων

sfcl:Expression (language version)

Identifier	Range	Cardinality	Value type	Description
dcterms:abstract	*rdfs:Literal*	1	Literal	abstract of the judgment
dcterms:language	• *fi* • *sv*	1	Literal	language
dcterms:title	*rdfs:Literal*	1	Literal	name of the document
sfcl:isEmbodiedBy	*sfcl:Manifestation*	3	URI	content format
sfcl:realizes	*sfcl:Judgement*	1	URI	judgment

Εικόνα 34: Οι κλάσεις και οι ιδιότητες του περιεχομένου των δικαστικών αποφάσεων

sfcl:Format (content format)

Identifier	Range	Cardinality	Value type	Description
sfcl:embodies	*sfcl:Expression*	1	URI	language version
sfcl:text	*rdfs:Literal*	1	Literal	text content
sfcl:html	*rdfs:Literal*	1	Literal	HTML content
sfcl:xml	*rdf:XMLLiteral*	1	XMLLiteral	original XML

Μελετώντας τις παραπάνω κλάσεις [εικόνες 32-34] και ιδιότητες καταλήξαμε στη σύνταξη του παρακάτω ερωτήματος:

PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

PREFIX dcterms: <http://purl.org/dc/terms/>

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX sfcl: <http://data.finlex.fi/schema/sfcl/>

PREFIX sfl: <http://data.finlex.fi/schema/sfl/>

PREFIX eli: <http://data.europa.eu/eli/ontology#>

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>

SELECT ?caselaw ?judgement_date ?issued_date ?desc_label ?label_subj ?language
?time_difference
WHERE {
  ?caselaw rdf:type sfcl:Judgment;
        dcterms:date ?judgement_date;
        dcterms:description ?description;
        dcterms:issued   ?issued_date;
        dcterms:subject ?subject;
        sfcl:isRealizedBy ?expression.
  ?expression dcterms:language ?language.
  ?subject skos:prefLabel ?label_subj.
  ?description skos:prefLabel ?desc_label.

FILTER ((?language = "fi") && (lang(?desc_label) = "fi"))
FILTER (?judgement_date > "2008-01-01"^^xsd:date)

BIND( fn:days-from-duration(?issued_date - ?judgement_date)/30 as ?time_difference)
   }
ORDER BY DESC(?time_difference)
LIMIT 10
```

Το ερώτημα αναζητά τις δικαστικές αποφάσεις (?judgments), την ημερομηνία εκδίκασης (?judgment_date), την ημερομηνία έκδοσης της απόφασης (?issued_date), το χρονικό διάστημα που μεσολαβεί μεταξύ της εκδίκασης μιας υπόθεσης και της έκδοσης απόφασης (?time_difference), τις λέξεις – κλειδιά της απόφασης (?desc_label, ?label_subject) και τη γλώσσα της απόφασης (?language). Και εδώ απαραίτητη προϋπόθεση είναι ο ορισμός prefixes πχ sfcl κλπ. Ζητήσαμε να επιστραφούν τα 10 πρώτα αποτελέσματα που έχουν εκδοθεί μετά από την 01.01.2008 και έχουν το θέμα και τις λέξεις – κλειδιά στα φιλανδικά καθώς υπάρχει η αντίστοιχη εκδοχή τους και στα σουηδικά.

Παρακάτω [εικόνα 35] είναι το ερώτημα στο γραφικό περιβάλλον του yasgui και στην εικόνα 36 παρουσιάζονται τα αποτελέσματά της εκτέλεσής του.

Εικόνα 35: Το ερώτημα για τις δικαστικές αποφάσεις

http://data.finlex.fi/sparql

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sfcl: <http://data.finlex.fi/schema/sfcl/>
PREFIX sfl: <http://data.finlex.fi/schema/sfl/>
PREFIX eli: <http://data.europa.eu/eli/ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>

SELECT ?caselaw ?judgement_date ?issued_date ?desc_label ?label_subj ?language ?time_difference
WHERE {
  ?caselaw rdf:type sfcl:Judgment;
           dcterms:date ?judgement_date;
           dcterms:description ?description;
           dcterms:issued   ?issued_date;
           dcterms:subject ?subject;
           sfcl:isRealizedBy ?expression.
  ?expression dcterms:language ?language.
  ?subject skos:prefLabel ?label_subj.
  ?description skos:prefLabel ?desc_label.

FILTER ((?language = "fi") && (lang(?desc_label) = "fi"))
FILTER (?judgement_date > "2008-01-01"^^xsd:date)
BIND( fn:days-from-duration(?issued_date - ?judgement_date)/30 as ?time_difference)
   }
ORDER BY DESC(?time_difference)
LIMIT 10
```

Εικόνα 36: Αποτέλεσμα ερωτήματος με φθίνουσα σειρά ως προς το χρονικό διάστημα μεταξύ εκδίκασης υπόθεσης – έκδοσης δικαστικής απόφασης

	caselaw	judgement_date	issued_date	desc_label	label_subj	language	time_difference
1	http://data.finlex.fi/ecli/kko/2013/50	"2103-01-16"^^xsd:date	"2013-06-26"^^xsd:date	"Laiton uhkaus"@fi	"Siviiliasia"@fi	fi	"1090.333333333333333333333333"^^xsd:decimal
2	http://data.finlex.fi/ecli/kko/2013/50	"2103-01-16"^^xsd:date	"2013-06-26"^^xsd:date	"Sananvapaus"@fi	"Siviiliasia"@fi	fi	"1090.333333333333333333333333"^^xsd:decimal
3	http://data.finlex.fi/ecli/kko/2013/50	"2103-01-16"^^xsd:date	"2013-06-26"^^xsd:date	"Syyteoikeus"@fi	"Siviiliasia"@fi	fi	"1090.333333333333333333333333"^^xsd:decimal
4	http://data.finlex.fi/ecli/kko/2013/69	"2011-09-20"^^xsd:date	"2013-09-23"^^xsd:date	"Kunnianloukkaus"@fi	"Siviiliasia"@fi	fi	"24.466666666666666666666666"^^xsd:decimal
5	http://data.finlex.fi/ecli/kko/2013/69	"2011-09-20"^^xsd:date	"2013-09-23"^^xsd:date	"Sananvapaus"@fi	"Siviiliasia"@fi	fi	"24.466666666666666666666666"^^xsd:decimal
6	http://data.finlex.fi/ecli/kko/2013/69	"2011-09-20"^^xsd:date	"2013-09-23"^^xsd:date	"Törkeä kunnianloukkaus"@fi	"Siviiliasia"@fi	fi	"24.466666666666666666666666"^^xsd:decimal
7	http://data.finlex.fi/ecli/kko/2013/69	"2011-09-20"^^xsd:date	"2013-09-23"^^xsd:date	"Yksityiselämän suoja"@fi	"Siviiliasia"@fi	fi	"24.466666666666666666666666"^^xsd:decimal
8	http://data.finlex.fi/ecli/kko/2013/70	"2011-09-20"^^xsd:date	"2013-09-23"^^xsd:date	"Kunnianloukkaus"@fi	"Siviiliasia"@fi	fi	"24.466666666666666666666666"^^xsd:decimal
9	http://data.finlex.fi/ecli/kko/2013/70	"2011-09-20"^^xsd:date	"2013-09-23"^^xsd:date	"Sananvapaus"@fi	"Siviiliasia"@fi	fi	"24.466666666666666666666666"^^xsd:decimal
10	http://data.finlex.fi/ecli/kko/2012/64	"2010-07-01"^^xsd:date	"2012-06-25"^^xsd:date	"Keskinäinen kiinteistöosakeyhtiö"@fi	"Rikosasia"@fi	fi	"24.166666666666666666666666"^^xsd:decimal

Showing 1 to 10 of 10 entries (in 0.901 seconds)

Κάνοντας μετάφραση της ετικέτας (Φινλαδικά -> Αγγλικά) της απόφασης διαπιστώνουμε ότι για τις υποθέσεις που σχετίζονται με την ‘Ελευθερία του λόγου’ και

την ‘δυσφήμιση’, απαιτείται μεγάλο χρονικό διάστημα μέχρι την έκδοση της απόφασης. Με την χρήση της μεταβλητής που υπολογίζει τους μήνες που μεσολαβούν από την συζήτηση της υπόθεσης μέχρι την έκδοση κάθε απόφασης, μπορούμε επίσης να εξάγουμε πολύ χρήσιμα συμπεράσματα σχετικά με την ταχύτητα απονομής της δικαιοσύνης, την απόδοση των δικαστικών λειτουργών και το είδος των υποθέσεων στις οποίες απαιτείται περισσότερος χρόνος για την έκδοση απόφασης. Οι πληροφορίες αυτές μπορούν να αξιοποιηθούν από τους νομικούς, από απλούς πολίτες αλλά και από τις κυβερνήσεις για την χάραξη πολιτικής. Είναι αυτονόητο ότι εάν υπήρχαν σωστά δομημένες οι αντίστοιχες πληροφορίες και σε άλλες χώρες, θα μπορούσε να γίνει μια πολύ ενδιαφέρουσα στατιστική και συγκριτική μελέτη για τον μέσο όρο έκδοσης των αποφάσεων πχ με οικονομικό αντικείμενο σε διάφορες χώρες και κατά πόσο υπάρχει συσχέτιση των δεδομένων που προκύπτουν με το ύψος των επενδύσεων μιας χώρας.

4.3 Η διαδικασία εισαγωγής του data set του statista.com στο Rapidminer και ανάλυση των αποτελεσμάτων

Για να μπορέσουμε να εκμεταλλευτούμε τις συνδέσεις μεταξύ των ανοιχτών συνδεδεμένων δεδομένων και να παρουσιάσουμε τις δυνατότητες ανάλυσης αυτών μέσω του εργαλείου RapidMiner LOD extension, προχωρήσαμε στις ακόλουθες ενέργειες:

• Εγγραφή και free trial member της ιστοσελίδας https://www.statista.com/chart/15584/the-number-of-tourist-arrivals-and-total-inbound-expenditure/ (βλ. εικόνα 37).

• Περιήγηση σε ενδιαφέροντα στατιστικά datasets και επιλογή του dataset που αναφέρει τις επισκέψεις τουριστών (σε εκατομμύρια) και τα χρήματα που δαπάνησαν (σε δισεκατομμύρια δολάρια) ανά χώρα κατά το έτος 2017.

Εικόνα 37: Διάγραμμα του σετ δεδομένων από statista.com

• Αποθήκευση του αρχείου ως CSV [εικόνα 38] .

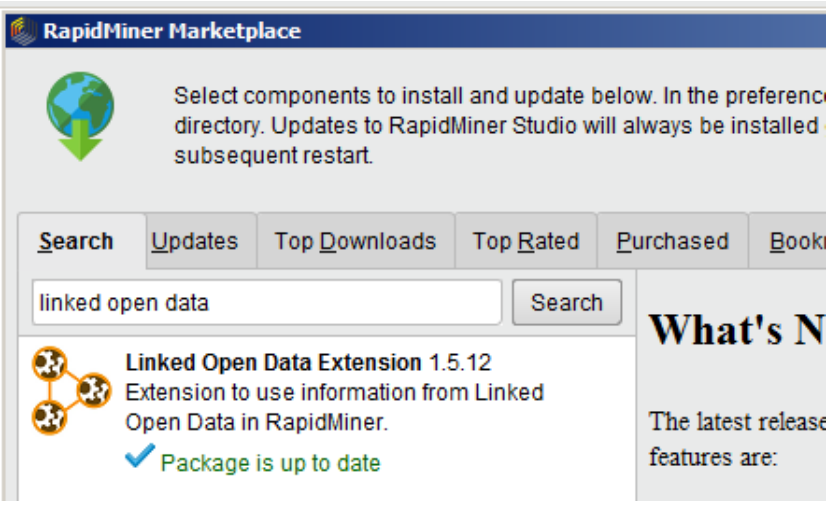
Εικόνα 38: Προεπισκόπηση των δεδομένων μας (σε μορφή csv)

	A	B	C	D	E	F
1	Country,Expenditure(in billion US dollars),Arrivals(in million)					
2	United States,201.70,75.90					
3	Spain,68.00,81.80					
4	France,60.70,86.90					
5	Thailand,57.50,35.40					
6	United Kingdom,51.20,37.70					
7	Italy,44.20,58.30					
8	Australia,41.70,8.80					
9	Germany,39.80,37.50					
10	China,35.60,17.30					
11	Japan,34.10,28.70					

Το εργαλείο που θα χρησιμοποιήσουμε για την εισαγωγή και ανάλυση των δεδομένων μας είναι το Rapidminer. Μετά την εγκατάσταση της έκδοσης Rapidminer Studio 7.1, το επόμενο βήμα είναι η αναζήτηση της επέκτασης του Linked Open Data Extension από τον χώρο της Αγοράς Επεκτάσεων του Rapidminer [εικόνα 39]. Εντοπίζουμε την

επέκταση, επιβεβαιώνουμε ότι πρόκειται για την έκδοση 1.5.12 και προχωράμε με την εγκατάστασή της.

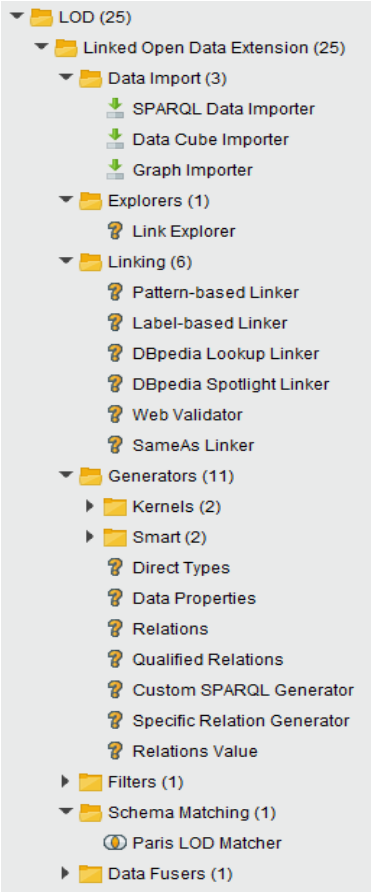
Εικόνα 39: Εντοπισμός της επέκτασης Linked Open Data Extension στο Market place του Rapidminer



Αμέσως μετά την ολοκλήρωση της εγκατάστασης του απαραίτητου εργαλείου για την επεξεργασία και ανάλυση των Ανοιχτών Συνδεδεμένων Δεδομένων, περιηγούμαστε στους Operators που διαθέτει [εικόνα 40]. Εντοπίζουμε τις γενικότερες κατηγορίες αυτών οι οποίες είναι:

- Εισαγωγή δεδομένων (Data Import)
- Εξερευνητές (Explorer)
- Διασύνδεση (Linking)
- Γεννήτριες (Generators)
- Φίλτρα (Filters)
- Ανάμειξη δεδομένων (Data fusers)
- Αναπαράσταση (Representation)

Εικόνα 40: Εξερεύνηση των Operators της επέκτασης Συνδεδεμένων Ανοιχτών Δεδομένων (LOD extension)



Έχοντας ολοκληρώσει επιτυχώς τα παραπάνω βήματα (αποθήκευση του στατιστικού σετ δεδομένων στην κατάλληλη μορφή Countries_Tourists_expenditures_20180926.csv, εγκατάσταση της εφαρμογής και της επέκτασης του Rapidminer) θα ξεκινήσουμε τη διαδικασία ανάλυσης των δεδομένων μας. Η μεθοδολογία που ακολουθήσαμε περιλαμβάνει 3 στάδια, τα οποία είναι τα εξής:

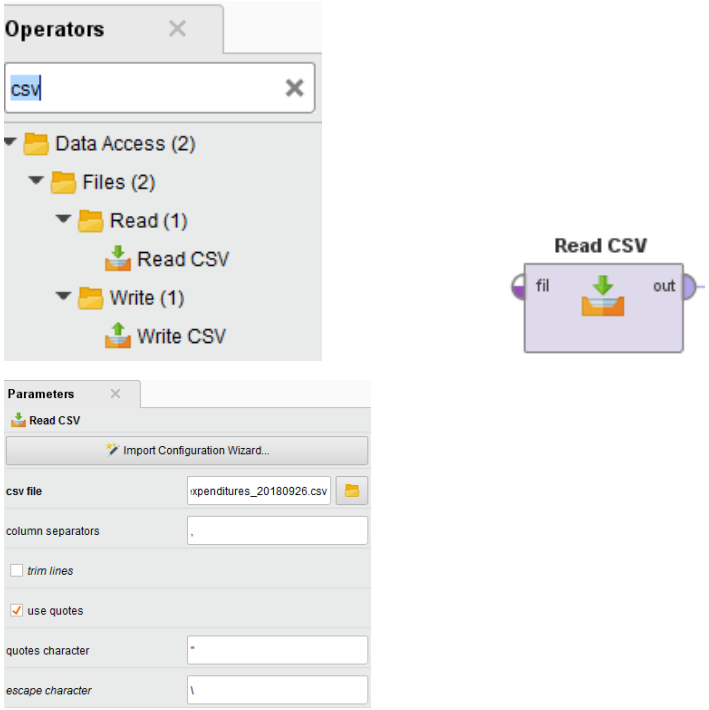
1. Εισαγωγή και σύνδεση των δεδομένων μας με την Dbpedia.
2. Προσθήκη επιπλέον πληροφορίας στο αρχικό σετ δεδομένων.
3. Επιλογή των μεταβλητών και Ανάλυση Συσχέτισης μεταξύ των μεταβλητών.

Ακολουθεί αναλυτική καταγραφή των βημάτων με λεκτικές περιγραφές και στιγμιότυπα οθόνης.

1ο στάδιο - Εισαγωγή και σύνδεση των δεδομένων μας με την Dbpedia

Κάνουμε αναζήτηση του κατάλληλου Operator (read csv) στο πλαίσιο διαλόγου και τον εισάγουμε στον καμβά της διαδικασίας (process) [βλ. εικόνα 41].

Εικόνα 41: Read_CSV Operator – Ορισμός παραμέτρων



Μετά το Read CSV operator και προκειμένου να ολοκληρώσουμε την εισαγωγή των δεδομένων μας, έχουμε 2 επιλογές:

- Είτε μέσω του Οδηγού Εισαγωγής (Import Configuration Wizard)
- Είτε ορίζοντας απευθείας την τοποθεσία και το όνομα του αρχείου, την παράμετρο διαχωρισμού των πεδίων (column separators) καθώς και την μορφή (format) και κωδικοποίηση των δεδομένων μας.

Έπειτα, θα πρέπει να ορίσουμε το πεδίο το οποίο διαδραματίζει κύριο ρόλο στο σύνολο της διαδικασίας, το οποίο στη δική μας περίπτωση θα είναι το πεδίο 'Country' με target role = regular. Η προηγούμενη διαδικασία θα υλοποιηθεί με τον operator 'Set Role' [εικόνα 42].

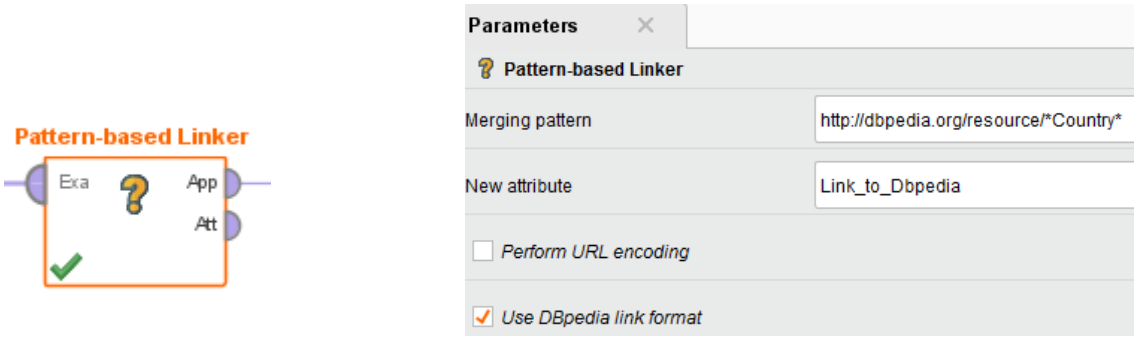
Εικόνα 42: Set Role Operator – Ορισμός παραμέτρων

Επόμενος στόχος είναι η διασύνδεση των τιμών του πεδίου ‘Country’ με κάποιο endpoint και συγκεκριμένα με την Dbpedia. Για να το πετύχουμε αυτό, θα πρέπει να επιλέξουμε κάποιον Operator από την Κατηγορία της Διασύνδεσης (Linking). Επιλέγουμε τον **Pattern-Based Linker**, ο οποίος δημιουργεί συνδέσμους χρησιμοποιώντας ένα καθορισμένο μοτίβο URI. Ο συγκεκριμένος operator απαιτεί ως παραμέτρους: (α) το σχέδιο σύνδεσης (link pattern), (β) το όνομα του νέου χαρακτηριστικού και (γ) προαιρετικά την εκτέλεση κωδικοποίησης URL (με αντικατάσταση των ειδικών χαρακτήρων και τη δημιουργία κατάλληλων συνδέσμων UTF-8) ή μιας συγκεκριμένης σύνδεσης για την DBpedia. Ο Pattern-Based Linker Operator θα προσπαθήσει να διαμορφώσει, σύμφωνα με τις παραπάνω παραμέτρους, το Uri για κάθε μια από τις τιμές του πεδίου ‘Country’ (Ηνωμένες Πολιτείες, Ισπανία κλπ) στη βάση δεδομένων της Dbpedia.

Τα βήματα που θα ακολουθήσουμε με τον Pattern-Based Linker είναι τα εξής:

- Από την κατηγορία της Διασύνδεσης (Linking) επιλέγουμε τον Operator με την ονομασία ‘Pattern-based Linker’ [εικόνα 43] με την χρήση του οποίου θα προσθέσουμε στο σετ δεδομένων μας ένα επιπλέον πεδίο που θα συγχωνεύει: α) το μοντέλο που ορίζουμε στην παράμετρο ‘Merging Pattern’ μαζί με β) το πεδίο που θα ορίσουμε ανάμεσα στους αστερίσκους (*Attribute*).
- Το επόμενο βήμα είναι να ορίσουμε την παράμετρο ‘Merging pattern’ -> http://dbpedia.org/resource/*Country* έτσι ώστε η συγχώνευση να πραγματοποιηθεί για κάθε μια τιμή του πεδίου ‘Country’.
- Αμέσως μετά, ορίζουμε το label (ετικέτα) του νέου attribute που θα προστεθεί. Συγκεκριμένα: -> Link_to_Dbpedia
- Τέλος, «τσεκάρουμε» την επιλογή ‘Use Dbpedia link format’ προκειμένου να γίνει χρήση του μοντέλου ονομασίας της Dbpedia. Ειδικότερα θα γίνει αντικατάσταση του χαρακτήρα ‘space’ ανάμεσα στις λέξεις με “underscore” (όπου υπάρχει). Για παράδειγμα United Kingdom -> United_Kingdom.

Εικόνα 43: Pattern-based Linked Operator – Ορισμός παραμέτρων

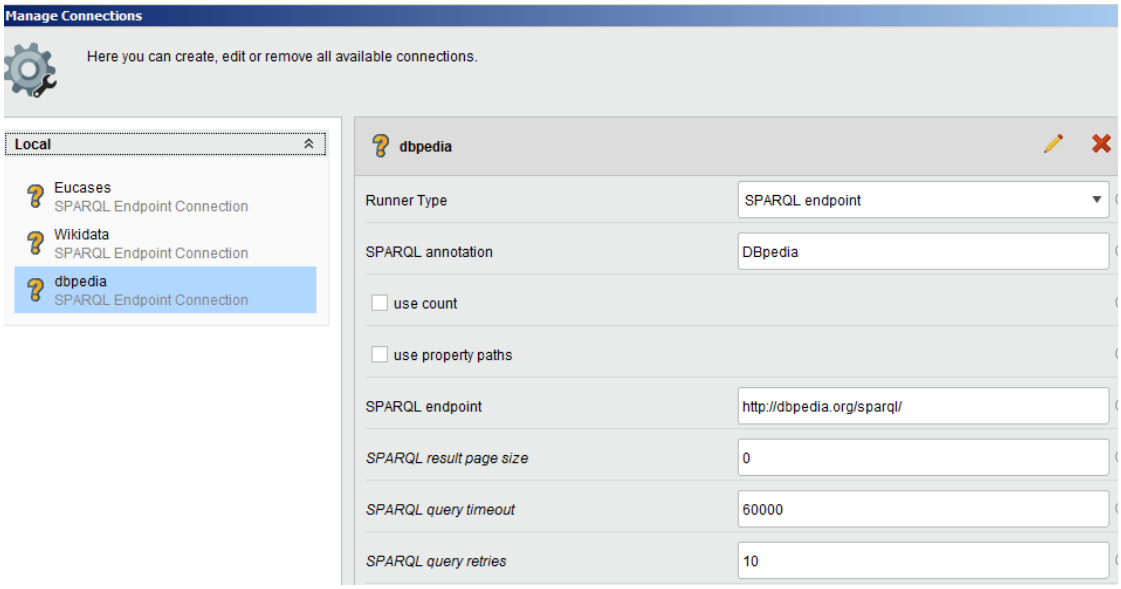


Δεδομένου ότι μερικοί linking operators (όπως ο Pattern-based τον οποίον χρησιμοποιήσαμε) μπορεί να δημιουργήσουν συνδέσμους που στην πραγματικότητα δεν υπάρχουν, ο Επαληθευτής ιστού (Web Validator) του RapidMiner ελέγχει την ύπαρξη κάθε συνδέσμου (URI) και αφαιρεί τα στιγμιότυπα για τα οποία δεν υπάρχει ή δεν βρέθηκε κάποιος σύνδεσμος.

Συνεπώς, θέλοντας να εφαρμόσουμε το παραπάνω μοντέλο ελέγχου, επιλέγουμε τον operator **Web Validator** [εικόνα 45] από την κατηγορία των Linkers.

Πριν συνεχίσουμε την διαδικασία ορισμού παραμέτρων για τον Web Validator θα πρέπει να δημιουργήσουμε την σύνδεση με το endpoint της Dbpedia. Από το κεντρικό μενού του Rapidminer, επιλέγουμε Συνδέσεις (Connections) και μετά Διαχείριση Συνδέσεων (Manage Connections). Κάνουμε προσθήκη Νέας Σύνδεσης με την ονομασία ‘dbpedia’ και ορίζουμε ως Runner Type , την τιμή Sparql Endpoint. Έπειτα, ορίζουμε το url του endpoint ως http://dbpedia.org/sparql [βλ. εικόνα 44].

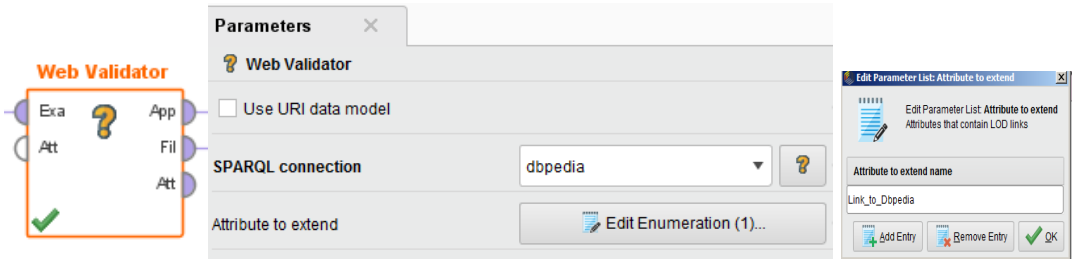
Εικόνα 44: Προσθήκη νέας σύνδεσης Sparql Endpoint (Dbpedia)



Η διαδικασία με τον Web Validator είναι η εξής:

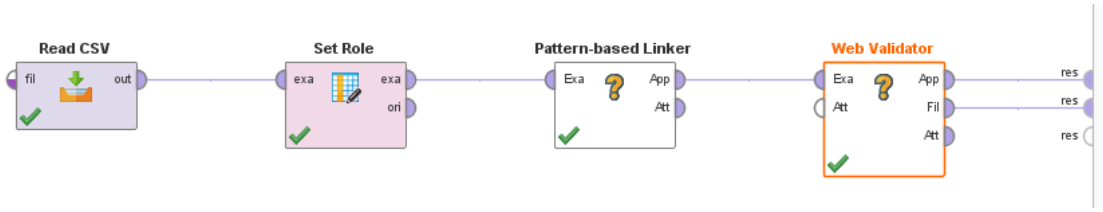
- Επιλέγουμε το Sparql endpoint στο οποίο θα γίνει ο έλεγχος εγκυρότητας των URI's. Στη δική μας περίπτωση επιλέγουμε την dbpedia, την οποία δημιουργήσαμε στο προηγούμενο βήμα.
- Επίσης ορίζουμε ως 'Attribute to extend' την τιμή του πεδίου που θέλουμε να επεκτείνουμε, δηλαδή την 'Link_to_Dbpedia'.

Εικόνα 45: Web Validator (μέσω Dbpedia)



Προκειμένου να ελέγξουμε την ορθή εκτέλεση και να επιβεβαιώσουμε όσα έχουμε αναφέρει παραπάνω, συνδέουμε την έξοδο του Appended Set με την είσοδο των αποτελεσμάτων (res) και την έξοδο του Filtered Set (μετά την τυχόν αφαίρεση συνδέσμων που δεν είναι έγκυροι ή δεν υπάρχουν) [βλ. εικόνα 46].

Εικόνα 46: Αναπαράσταση διαδικασίας (Read CSV-Linking-Web Validation)



Εκτελώντας  τη διαδικασία, παράγονται τα εξής αποτελέσματα [εικόνα 47, 48]:

Εικόνα 47: Αποτελέσματα διαδικασίας (Web Validator-Appended)

ExampleSet (Web Validator) | **ExampleSet (Pattern-based Linker)**

ExampleSet (10 examples, 0 special attributes, 5 regular attributes)

Row No.	Expenditure...	Arrivals(in ...	Country	Link_to_Dbpedia	RecordExist...
1	201.700	75.900	United States	http://dbpedia.org/resource/United_States	true
2	68	81.800	Spain	http://dbpedia.org/resource/Spain	true
3	60.700	86.900	France	http://dbpedia.org/resource/France	true
4	57.500	35.400	Thailand	http://dbpedia.org/resource/Thailand	true
5	51.200	37.700	United Kingd...	http://dbpedia.org/resource/United_Kingdom	true
6	44.200	58.300	Italy	http://dbpedia.org/resource/Italy	true
7	41.700	8.800	Australia	http://dbpedia.org/resource/Australia	true
8	39.800	37.500	Germany	http://dbpedia.org/resource/Germany	true
9	35.600	17.300	China	http://dbpedia.org/resource/China	true
10	34.100	28.700	Japan	http://dbpedia.org/resource/Japan	true

Εικόνα 48: Αποτελέσματα διαδικασίας (Web Validator - Filtered)

ExampleSet (Web Validator) | ExampleSet (Pattern-based Linker)

ExampleSet (10 examples, 0 special attributes, 4 regular attributes)

Row No.	Expenditure...	Arrivals(in ...	Country	Link_to_Dbpedia
1	201.700	75.900	United States	http://dbpedia.org/resource/United_States
2	68	81.800	Spain	http://dbpedia.org/resource/Spain
3	60.700	86.900	France	http://dbpedia.org/resource/France
4	57.500	35.400	Thailand	http://dbpedia.org/resource/Thailand
5	51.200	37.700	United Kingd...	http://dbpedia.org/resource/United_Kingdom
6	44.200	58.300	Italy	http://dbpedia.org/resource/Italy
7	41.700	8.800	Australia	http://dbpedia.org/resource/Australia
8	39.800	37.500	Germany	http://dbpedia.org/resource/Germany
9	35.600	17.300	China	http://dbpedia.org/resource/China
10	34.100	28.700	Japan	http://dbpedia.org/resource/Japan

Όπως παρατηρούμε στην εικόνα 47 (Appended Set) έχει προστεθεί η νέα στήλη του συνδέσμου (Link_to_Dbpedia) καθώς και μια επιπλεόν στήλη, η οποία αναφέρει την εγκυρότητα (true) ή όχι (false) των URI's στο endpoint της Dbpedia. Σε περίπτωση που είχαμε κάποια εσφαλμένο URI ή ανύπαρκτο, τότε η συγκεκριμένη γραμμή (Row) θα είχε την τιμή 'False' στην τελευταία στήλη, όπως επίσης θα είχε αφαιρεθεί από τα αποτελέσματα της εικόνας 48 (Filtered Set).

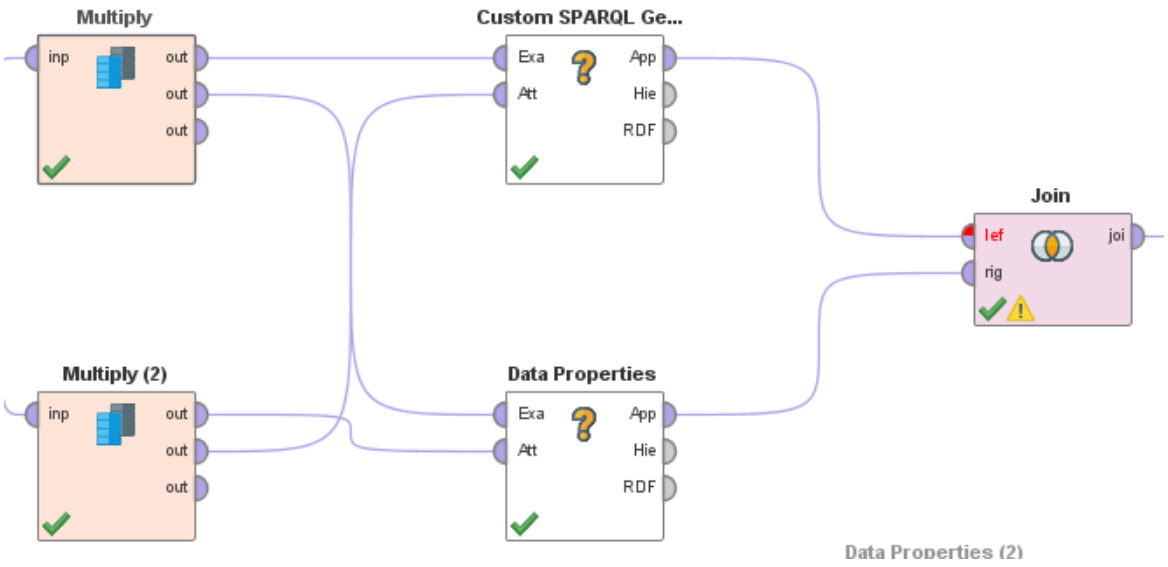
2ο στάδιο - Προσθήκη επιπλέον πληροφορίας στο αρχικό σετ δεδομένων

Προκειμένου να προστεθούν πληροφορίες από πολλές πηγές ή σχετικά με διαφορετικές οντότητες σε ένα dataset, τα δεδομένα μπορούν να πολλαπλασιαστούν με τον Multiply Operator του RapidMiner (στη ενότητα "Process control"). Στην περίπτωση αυτή η έξοδος και των δύο συνδέσμων ενώνεται με τον Join operator (στην ενότητα "Data Transformation"). Για να εκτελεστεί επιτυχώς το Join, τα αρχικά δεδομένα πρέπει να ορίζουν τουλάχιστον μία στήλη αναγνωριστικού.

Σύμφωνα με τα παραπάνω εισάγουμε δύο Multiply operators, στις εισόδους των οποίων συνδέουμε την έξοδο του Filtered Set και του Attributes Appended του Web Validator. Αυτό συμβαίνει καθώς οι γεννήτριες μεταβλητών (Attribute Generators) τις οποίες θα χρησιμοποιήσουμε είναι 2 και είναι οι εξής [εικόνα 49]:

- Ο **Custom SPARQL generator**. Σε περιπτώσεις που ο χρήστης γνωρίζει το σύνολο δεδομένων που θα χρησιμοποιήσει και θέλει να προσθέσει συγκεκριμένα δεδομένα που δεν καλύπτονται από κανένα από τις προεπιλεγμένους generators, υπάρχει η δυνατότητα ορισμού ατομικών δηλώσεων SPARQL που γίνεται μέσω του εν λόγω προσαρμοσμένου generator. Στις συγκεκριμένες δηλώσεις ο χρήστης μπορεί να χρησιμοποιήσει συνδέσμους που δημιουργούνται από linkers, οι οποίοι περικλείονται σε αστερίσκους.
- Ο **Data Properties generator**, ο οποίος δημιουργεί ένα attribute για κάθε κυριολεκτική αξία (literal value) που υπάρχει στα συνδεδεμένα στιγμιότυπα. Εκτελούνται επίσης μερικές βασικές εικασίες των τύπων δεδομένων π.χ., οι αριθμητικές τιμές σημειώνονται ως τέτοιες και το λοιπά αριθμητικά σύμβολα χωρίς ρητό τύπο αναλύονται σε αριθμούς, αν είναι δυνατόν.

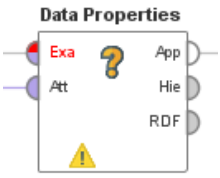
Εικόνα 49: Σύνδεση Multiply operators και Attribute Generators (Data Properties – Custom SPARQL Generator- Join)

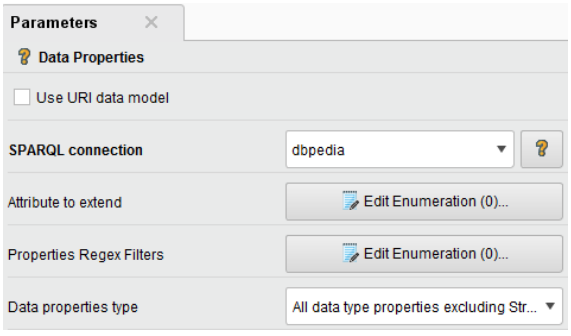


Στον operator **Data Properties** ορίζουμε τις εξής παραμέτρους [εικόνα 50]:

- Επιλέγουμε ως σύνδεση SPARQL την dbpedia.
- Ως τύπο χαρακτηριστικών , επιλέγουμε 'All data type properties excluding Strings' προκειμένου να λάβουμε αριθμητικά στοιχεία (όχι αλφαριθμητικά).

Εικόνα 50: Data Properties (Attribute generator – Dbpedia)

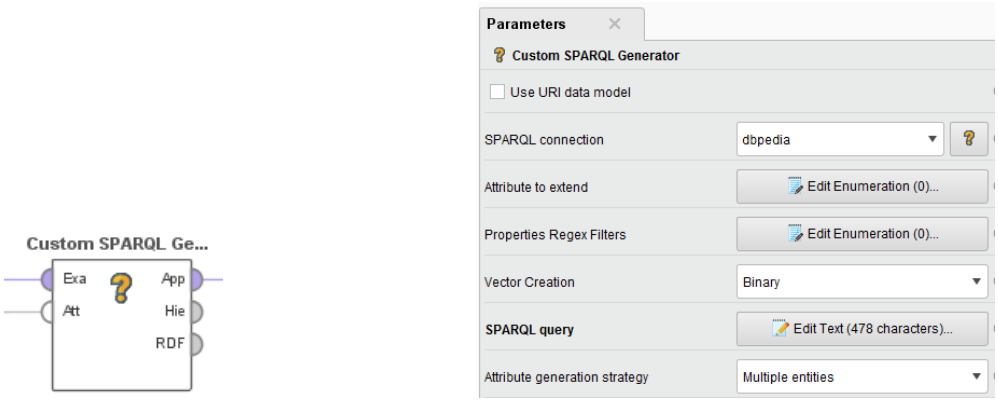




Στον operator **Custom SPARQL Generator** [εικόνα 51]:

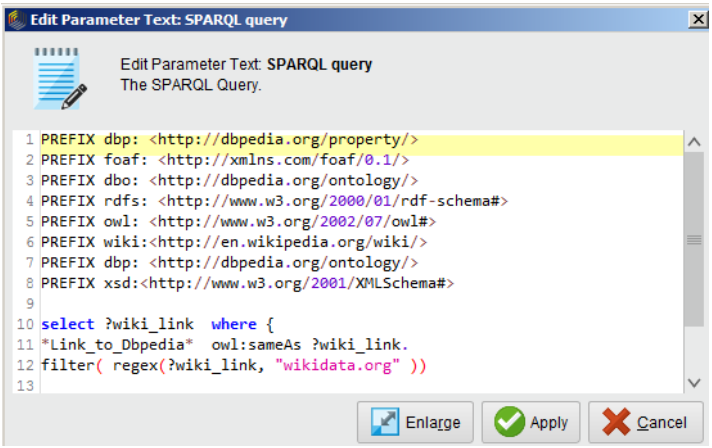
- Επιλέγουμε ως σύνδεση SPARQL την dbpedia.

Εικόνα 51: Custom SPARQL Generator (Attribute generator)



Απαραίτητη προϋπόθεση για την επιτυχή εκτέλεση του operator είναι η σύνταξη του ερωτήματος με τις ζητούμενες μεταβλητές [εικόνα 52].

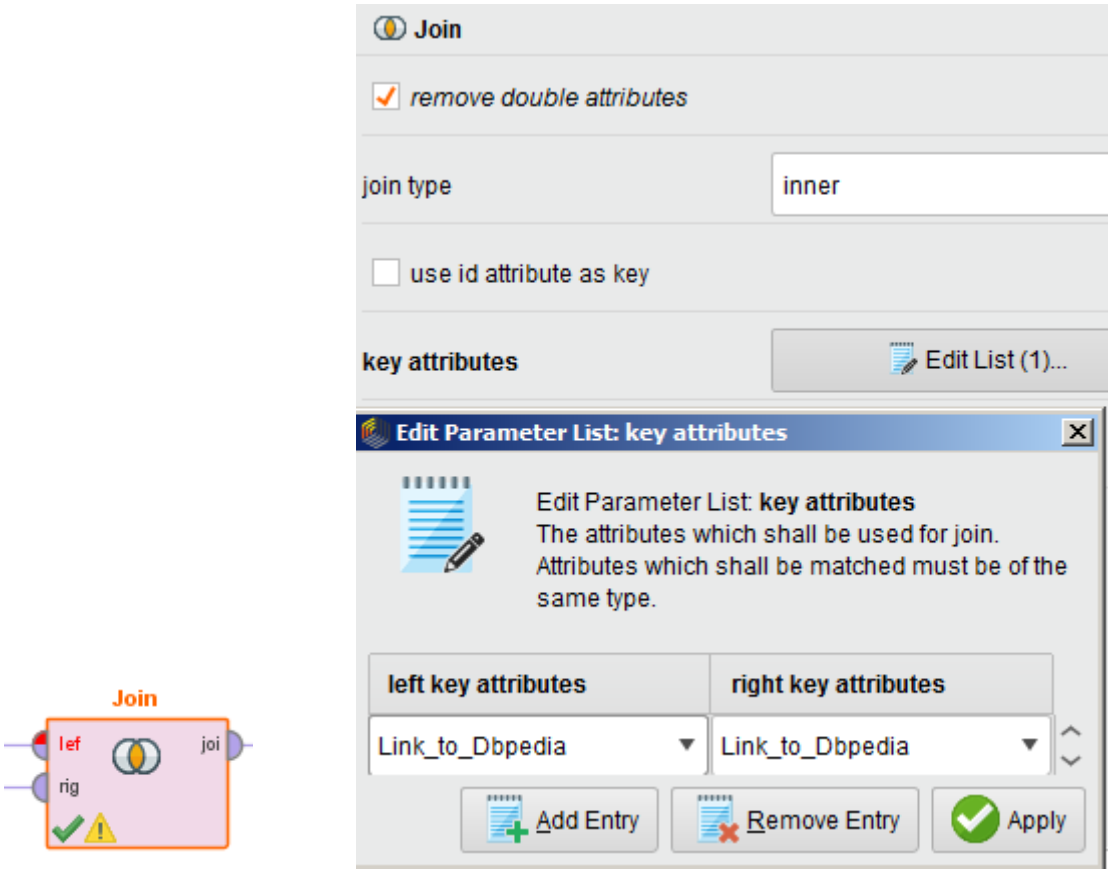
Εικόνα 52: SPARQL Ερώτημα (query – owl:sameAs)



Το ζητούμενο από το παραπάνω SPARQL ερώτημα είναι η άντληση του owl:sameAs uri για την κάθε μια χώρα από την Dbpedia. Απώτερος σκοπός μας είναι η διασύνδεση των ιδιοτήτων από την βιβλιοθήκη της Wikidata στο επόμενο στάδιο. Για τον λόγο αυτό περιορίσαμε τα αποτελέσματα του query ορίζοντας φίλτρο ως εξής: (filter(regex (?wiki_link, "wikidata.org")).

Επισημαίνεται επίσης, όπως αναφέραμε παραπάνω στην εισαγωγή του 2^{ου} σταδίου, ότι για να εξαχθούν επιτυχώς τα ζητούμενα αποτελέσματα θα πρέπει οι 2 έξοδοι των operators να συνδεθούν με έναν Join operator [εικόνα 53] ως ακολούθως (αναγνωριστικό πεδίο = Link_to_Dbpedia).

Εικόνα 53: Join Operator



Εκτελούμε  και πάλι τη διαδικασία όπου παράγονται τα εξής αποτελέσματα [εικόνα 54].

Εικόνα 54: Αποτελέσματα (Join - Appended set – μετά από Attributes Generator (2))

ExampleSet (Join)

ExampleSet (10 examples, 0 special attributes, 189 regular attributes) Filter (1

Row No.	Expenditure...	Arrivals(in ...	Country	Link_to_Dbpedia	Link_to_Dbpedia_wiki_link
1	201.700	75.900	United States	http://dbpedia.org/resource/United_States	http://www.wikidata.org/entity/Q30
2	68	81.800	Spain	http://dbpedia.org/resource/Spain	http://www.wikidata.org/entity/Q29
3	60.700	86.900	France	http://dbpedia.org/resource/France	http://www.wikidata.org/entity/Q142
4	57.500	35.400	Thailand	http://dbpedia.org/resource/Thailand	http://www.wikidata.org/entity/Q869
5	51.200	37.700	United Kingdom	http://dbpedia.org/resource/United_Kingdom	http://www.wikidata.org/entity/Q145
6	44.200	58.300	Italy	http://dbpedia.org/resource/Italy	http://www.wikidata.org/entity/Q38
7	41.700	8.800	Australia	http://dbpedia.org/resource/Australia	http://www.wikidata.org/entity/Q408
8	39.800	37.500	Germany	http://dbpedia.org/resource/Germany	http://www.wikidata.org/entity/Q183
9	35.600	17.300	China	http://dbpedia.org/resource/China	http://www.wikidata.org/entity/Q148
10	34.100	28.700	Japan	http://dbpedia.org/resource/Japan	http://www.wikidata.org/entity/Q17

Όπως παρατηρούμε στην εικόνα 54 (Appended Set) έχει προστεθεί η νέα στήλη του URI των χωρών στην Wikidata (**Link_to_Dbpedia_wiki_link**) καθώς και το σύνολο των ιδιοτήτων για την κάθε χώρα από την Dbpedia (189 attributes).

Αυτό το οποίο πετύχαμε μέσω του Custom SPARQL Generator είναι η έμμεση σύνδεση των δεδομένων μας με την Wikidata. Για παράδειγμα:

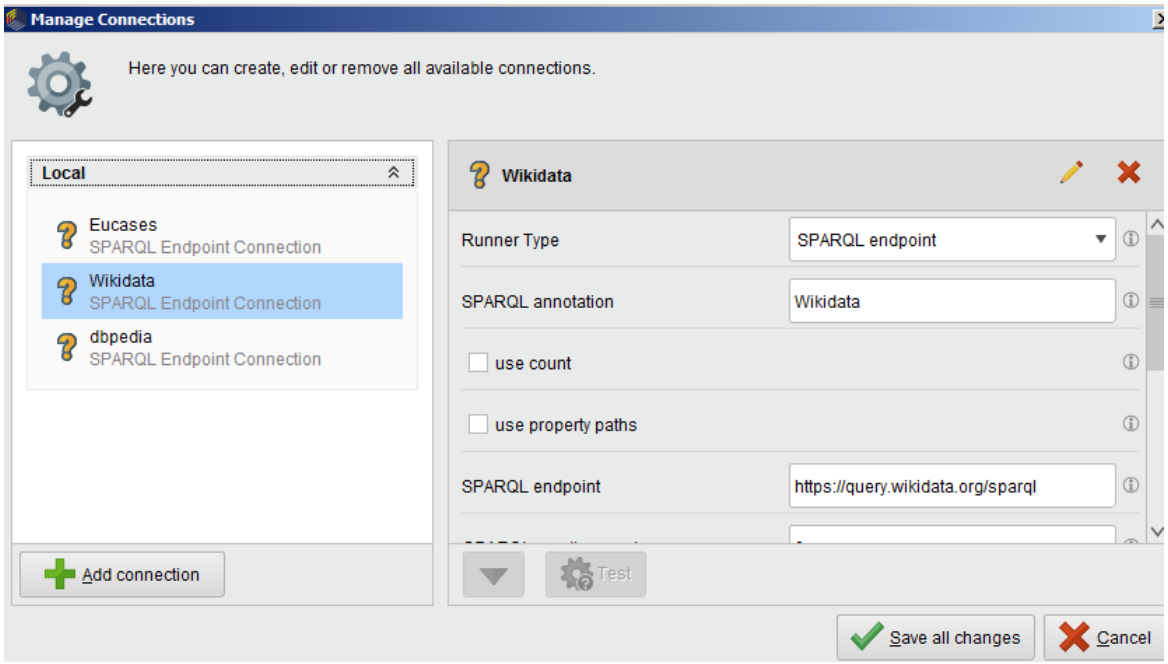
United States (Dbpedia): http://dbpedia.org/resource/United_States

United States (Wikidata): http://www.wikidata.org/entity/Q30

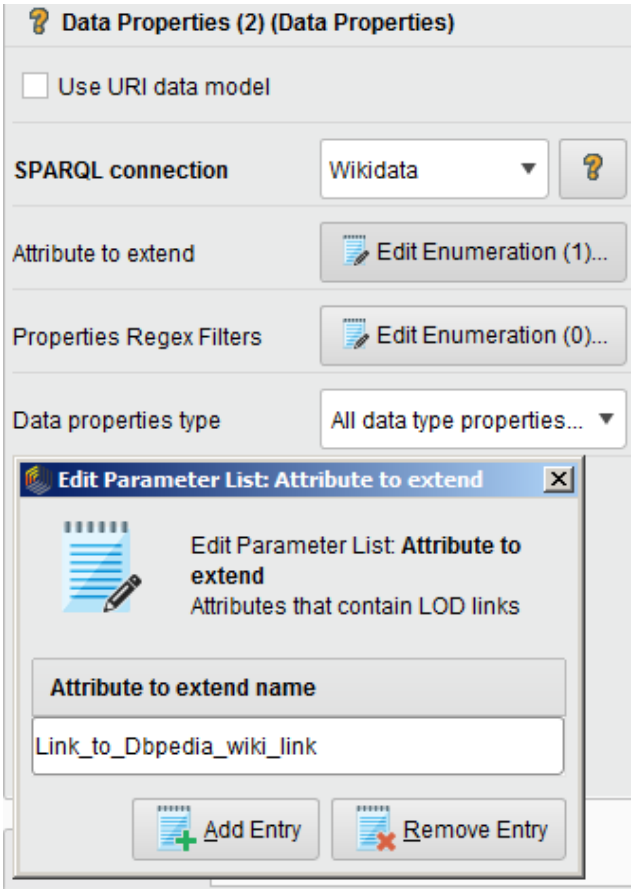
Παρατηρούμε ότι το μοτίβο ορισμού του URI μιας οντότητας διαφέρει ανάμεσα στην Dbpedia και στην Wikidata (βλ. United_States και Q30).

Ακολούθως θα εφαρμόσουμε εκ νέου τον operator Data Properties, με την μόνη διαφορά ότι αυτήν την φορά ως σύνδεση SPARQL θα οριστεί η Wikidata (https://query.wikidata.org/sparql) [εικόνα 55] και ως Attribute to extend η μεταβλητή Link_to_Dbpedia_wiki_link [εικόνα 56].

Εικόνα 55: Δημιουργία νέας σύνδεσης με Wikidata endpoint



Εικόνα 56: Data Properties Generator (2) - Wikidata



Εκτελούμε και πάλι τη διαδικασία όπου παράγονται τα εξής αποτελέσματα [εικόνα 57]:

Εικόνα 57: Αποτελέσματα (Data Properties – Wikidata)

ExampleSet (Data Properties (2))

ExampleSet (10 examples, 0 special attributes, 244 regular attributes)

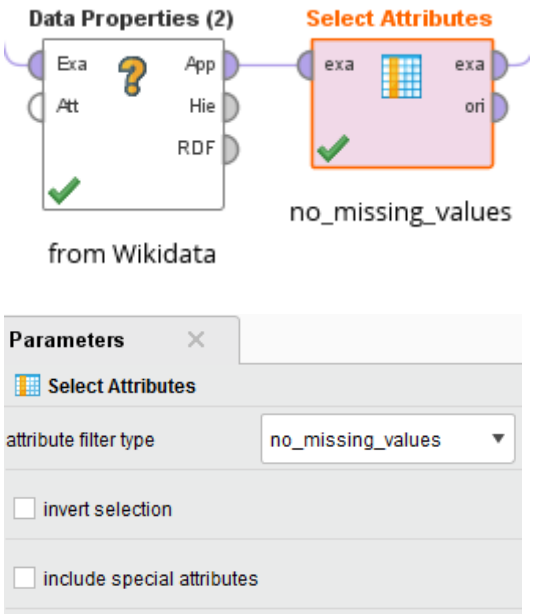
Row No.	Expenditure...	Arrivals(in ...	Country	Link_to_Dbpedia	Link_to_Dbpedia_wiki_link
1	201.700	75.900	United States	http://dbpedia.org/resource/United_States	http://www.wikidata.org/entity/Q30
2	68	81.800	Spain	http://dbpedia.org/resource/Spain	http://www.wikidata.org/entity/Q29
3	60.700	86.900	France	http://dbpedia.org/resource/France	http://www.wikidata.org/entity/Q142
4	57.500	35.400	Thailand	http://dbpedia.org/resource/Thailand	http://www.wikidata.org/entity/Q869
5	51.200	37.700	United Kingd...	http://dbpedia.org/resource/United_Kingdom	http://www.wikidata.org/entity/Q145
6	44.200	58.300	Italy	http://dbpedia.org/resource/Italy	http://www.wikidata.org/entity/Q38
7	41.700	8.800	Australia	http://dbpedia.org/resource/Australia	http://www.wikidata.org/entity/Q408
8	39.800	37.500	Germany	http://dbpedia.org/resource/Germany	http://www.wikidata.org/entity/Q183
9	35.600	17.300	China	http://dbpedia.org/resource/China	http://www.wikidata.org/entity/Q148
10	34.100	28.700	Japan	http://dbpedia.org/resource/Japan	http://www.wikidata.org/entity/Q17

Παρατηρούμε στην εικόνα 57 ότι τα πλήθος των μεταβλητών μετά από την σύνδεση και με την Wikidata έχει διαμορφωθεί στις 244 (55 επιπλεόν μεταβλητές συγκριτικά με το προηγούμενο αποτέλεσμα από την Dbpedia μόνο).

3ο στάδιο - Επιλογή των μεταβλητών και Ανάλυση Συσχέτισης μεταξύ των μεταβλητών

Έχοντας παρατηρήσει αρκετές «χαμένες τιμές» στα παραπάνω αποτελέσματα, θα χρησιμοποιήσουμε έναν operator, ο οποίος φιλτράρει τις ιδιότητες (Select Attributes). Ορίζουμε ως παράμετρο 'no_missing_values' και συνδέουμε την είσοδο του (Exa) με την έξοδο (Appended set) του προηγούμενου Data Properties (2) operator [εικόνα 58].

Εικόνα 58: Select Attributes (no_missing_values)



Εικόνα 59 : Αποτελέσματα (no missing values)

ExampleSet (Select Attributes)

ExampleSet (10 examples, 0 special attributes, 62 regular attributes) Filter (10 / 10 examples): all

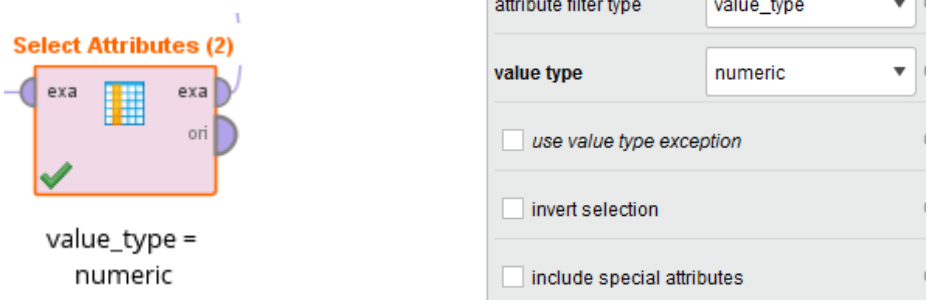
Row No.	Expenditure...	Arrivals(in ...	Country	Link_to_Dbpedia	Link_to_Dbpedia_wiki_link	Link_to_Dbp...	Link_to_Dbp
1	201.700	75.900	United States	http://dbpedia.org/resource/United_States	http://www.wikidata.org/entity/Q30	N	34.981
2	68	81.800	Spain	http://dbpedia.org/resource/Spain	http://www.wikidata.org/entity/Q29	N	92
3	60.700	86.900	France	http://dbpedia.org/resource/France	http://www.wikidata.org/entity/Q142	N	116
4	57.500	35.400	Thailand	http://dbpedia.org/resource/Thailand	http://www.wikidata.org/entity/Q869	N	132.047
5	51.200	37.700	United Kingd...	http://dbpedia.org/resource/United_Kingdom	http://www.wikidata.org/entity/Q145	N	255.561
6	44.200	58.300	Italy	http://dbpedia.org/resource/Italy	http://www.wikidata.org/entity/Q38	N	201.300
7	41.700	8.800	Australia	http://dbpedia.org/resource/Australia	http://www.wikidata.org/entity/Q408	S	2.800
8	39.800	37.500	Germany	http://dbpedia.org/resource/Germany	http://www.wikidata.org/entity/Q183	N	225.098
9	35.600	17.300	China	http://dbpedia.org/resource/China	http://www.wikidata.org/entity/Q148	N	144.016
10	34.100	28.700	Japan	http://dbpedia.org/resource/Japan	http://www.wikidata.org/entity/Q17	N	340.800

Παρατηρούμε στην εικόνα 59 ότι το πλήθος των μεταβλητών μετά το φιλτράρισμα και την αφαίρεση όσων περιείχαν «χαμμένες τιμές», έχει περιοριστεί στις 62.

Σημειώνεται βέβαια ότι αρκετές από αυτές είχαν την ίδια περιγραφή καθώς έχουμε αντλήσει δεδομένα από 2 διαφορετικές βιβλιοθήκες και είναι απόλυτα λογικό ότι κάποιες από τις ιδιότητες των χωρών ήταν κοινές (π.χ. ΑΕΠ, πληθυσμός, Γεωργαφικό μήκος και πλάτος κλπ).

Στην συνέχεια, προσθέτουμε τον ίδιο operator (Select Attributes) και ορίζουμε ως παράμετρο ‘value_type=numeric’ για να μας επιστρέψει μόνο τις αριθμητικές τιμές και συνδέουμε την είσοδο του (Exa) με την έξοδο (Exa) του προηγούμενου Select Attributes operator [εικόνα 60].

Εικόνα 60: Select Attributes 2 (value_type = numeric)



Εικόνα 61: Αποτελέσματα (value_type = numeric)

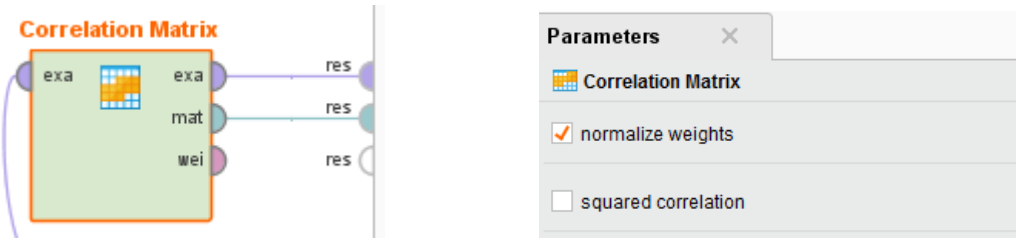
ExampleSet (10 examples, 0 special attributes, 44 regular attributes)

Row No.	Expenditure...	Arrivals(in ...	Link_to_Dbp...	Link_to_Dbp...	Link_to_Dbp...	Link_to_Dbp...	Link_to_Dbp...	Link_to_Dbp...	Link_to
1	201.700	75.900	34.981	0.915	40.800	745182619	57220	1	77
2	68	81.800	92	0.876	33.700	744504449	36143	42	3
3	60.700	86.900	116	0.888	30.100	744442908	41181	21.050	2
4	57.500	35.400	132.047	0.726	39.400	745296579	16706	29	100
5	51.200	37.700	255.561	0.907	31.600	745154135	42514	7	0
6	44.200	58.300	201.300	0.873	32.700	744125064	36191	29	12
7	41.700	8.800	2.800	0.935	33.600	745134554	47318	7.470	149
8	39.800	37.500	225.098	0.916	30.700	744929323	47033	23	13
9	35.600	17.300	144.016	0.727	46.200	744712446	15095	23	116
10	34.100	28.700	340.800	0.891	37.600	744601296	38731	46	139

Όπως φαίνεται στην εικόνα 61 το πλήθος των μεταβλητών μετά το φιλτράρισμα και την αφαίρεση των αλφαριθμητικών και λοιπών τιμών (πχ dates) έχει περιοριστεί στις 44.

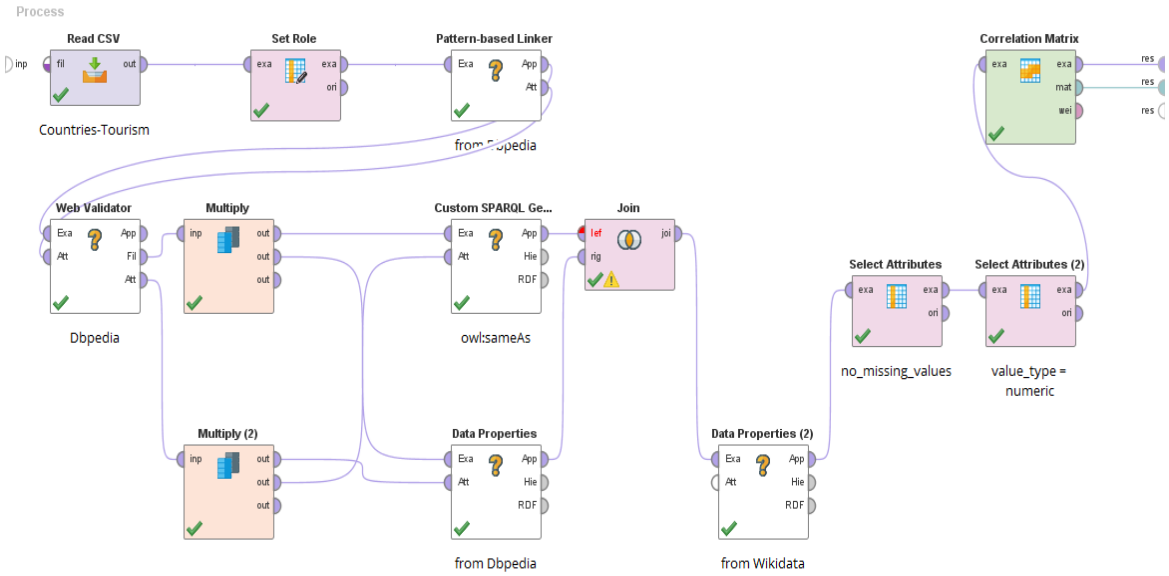
Ολοκληρώνοντας την έρευνά μας, χρησιμοποιούμε τον operator Correlation Matrix για να αναλύσουμε τις πιθανές συσχετίσεις ανάμεσα στις μεταβλητές του σετ δεδομένων μας σε σύνδεση πάντα με τον αριθμό των επισκεπτών και τα ποσά που ξόδεψαν το 2017 [εικόνα 62].

Εικόνα 62: Correlation Matrix



Εκτελούμε  τη διαδικασία και παρατηρούμε την επιτυχή ολοκήρωσή της χωρίς λάθη [εικόνα 63].

Εικόνα 63: Ολοκληρωμένη Διαδικασία (Επιτυχής εκτέλεση)



Εικόνα 64: Αποτελέσματα correlation που απεικονίζει τις συσχετίσεις (από τη μεγαλύτερη στη μικρότερη) των μεταβλητών συγκριτικά με τα ποσά που ξόδεψαν οι τουρίστες (expenditure)

Attributes	Expenditure(in billion US doll...
Link_to_Dbpedia_data_http://dbpedia.org/property/gdpNominal	0.751
Link_to_Dbpedia_wiki_link_data_http://wikiba.se/ontology#statements	0.687
Arrivals(in million)	0.523
Link_to_Dbpedia_data_http://dbpedia.org/ontology/areaTotal	0.517
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P4010	0.506
Link_to_Dbpedia_data_http://dbpedia.org/property/gdpPppPerCapita	0.500
Link_to_Dbpedia_data_http://dbpedia.org/property/gdpPpp	0.498
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P2299	0.479
Link_to_Dbpedia_data_http://dbpedia.org/ontology/PopulatedPlace/areaTotal	0.474
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P2046	0.474
Link_to_Dbpedia_data_http://dbpedia.org/property/populationDensityRank	0.460
Link_to_Dbpedia_data_http://dbpedia.org/property/gdpNominalPerCapita	0.455
Link_to_Dbpedia_wiki_link_data_http://wikiba.se/ontology#sitelinks	0.440
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P2132	0.415
Link_to_Dbpedia_wiki_link_data_http://schema.org/version	0.366
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P4841	0.360
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P1081	0.346

Εικόνα 65: Αποτελέσματα correlation που απεικονίζει τις συσχετίσεις (από τη μικρότερη στη μεγαλύτερη) των μεταβλητών συγκριτικά με τα ποσά που ξόδεψαν οι τουρίστες (expenditure)

Attributes	Expenditure(in billion US doll...
Link_to_Dbpedia_data_http://www.w3.org/2003/01/geo/wgs84_pos#long	-0.721
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P2884	-0.493
Link_to_Dbpedia_data_http://dbpedia.org/property/longm	-0.491
Link_to_Dbpedia_data_http://dbpedia.org/ontology/PopulatedPlace/populationDensity	-0.482
Link_to_Dbpedia_data_http://dbpedia.org/ontology/populationDensity	-0.482
Link_to_Dbpedia_data_http://dbpedia.org/property/areaRank	-0.434
Link_to_Dbpedia_data_http://dbpedia.org/property/hdiRank	-0.226
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P2131	-0.213

Εικόνα 66: Αποτελέσματα correlation που απεικονίζει τις συσχετίσεις (από τη μεγαλύτερη στη μικρότερη) των μεταβλητών συγκριτικά με τις αφίξεις των τουριστών

Attributes	Arrivals(in million) ↓
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P1198	0.635
Link_to_Dbpedia_data_http://www.w3.org/2003/01/geo/wgs84_pos#lat	0.542
Link_to_Dbpedia_wiki_link_data_http://wikiba.se/ontology#sitelinks	0.541
Expenditure(in billion US dollars)	0.523
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P1081	0.479
Link_to_Dbpedia_wiki_link_data_http://wikiba.se/ontology#statements	0.368
Link_to_Dbpedia_data_http://dbpedia.org/property/gdpPppPerCapita	0.317
Link_to_Dbpedia_wiki_link_data_http://wikiba.se/ontology#identifiers	0.311
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P2299	0.302

Εικόνα 67: Αποτελέσματα correlation που απεικονίζει τις συσχετίσεις (από τη μικρότερη στη μεγαλύτερη) των μεταβλητών συγκριτικά με τις αφίξεις των τουριστών

Attributes	Arrivals(in million) ↑
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P2134	-0.415
Link_to_Dbpedia_data_http://dbpedia.org/ontology/wikiPageRevisionID	-0.395
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P2219	-0.386
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P1279	-0.326
Link_to_Dbpedia_data_http://dbpedia.org/property/gini	-0.314
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P1082	-0.311
Link_to_Dbpedia_wiki_link_data_http://www.wikidata.org/prop/direct/P2046	-0.253
Link_to_Dbpedia_data_http://dbpedia.org/ontology/PopulatedPlace/areaTotal	-0.252
Link_to_Dbpedia_data_http://dbpedia.org/property/hdiRank	-0.246
Link_to_Dbpedia_data_http://dbpedia.org/property/gdpPppYear	-0.217
Link_to_Dbpedia_data_http://dbpedia.org/ontology/PopulatedPlace/populationDensity	-0.208

Από τα παραπάνω αποτελέσματα [εικόνες 64-67] του correlation matrix, συμπεραίνουμε ότι υπάρχει υψηλή συσχέτιση μεταξύ των χρημάτων που δαπάνησαν οι τουρίστες το 2017 με διάφορους οικονομικούς δείκτες των χωρών που επισκέφτηκαν όπως το ονομαστικό

ακαθάριστο εγχώριο προιόν των χωρών (gdp nominal), την αγοραστική δύναμη των κατοίκων (gdp ppp per capita), την ισοτιμία αγοραστικής δύναμης (gdp ppp) και φυσικά με τον αρθμό των αφίξεων. Γίνεται δηλαδή αντιληπτό ότι οι τουρίστες κατά την επιλογή του προορισμού τους λαμβάνουν υπόψη τους την εν γένει οικονομική κατάσταση της χώρας που πρόκεται να επισκεφτούν. Αντίθετα, παρατηρούμε ότι υπαρχει χαμηλή συσχέτιση με το γεωγραφικό μήκος (longm), την πληθυσμιακή πυκνότητα (population desnity) και την κατάταξη της χώρας από άποψη έκτασης (area rank). Αντίστοιχα, ο αριθμός των αφίξεων των τουριστών συνδέεται στενά με το ποσοστό ανεργίας (P1198) και το δείκτη ανθρώπινης ανάπτυξης (P1081) και λιγότερο στενά με τα αποθεματικά των χωρών (P2134), τον ρυθμό αναπτυξης του ονομαστικού ΑΕΠ (P2219) και με τον πληθωρισμό (P1279). Από τον συνδυασμό των παραπάνω εξάγεται σαν γενικό συμπέρασμα ότι η κατάσταση της οικονομίας μιας χώρας συσχετίζεται κατά βάση με το πόσα χρήματα θα δαπανήσουν οι τουρίστες χωρίς όμως να λειτουργεί αποτρεπτικά ως προς τις αφίξεις.

4.4 Συμπεράσματα που προέκυψαν από την ανάλυση των δεδομένων

Ανακεφαλαιώνοντας τα όσα εκθέσαμε στο παρόν κεφάλαιο, διαπιστώνουμε καταρχήν ότι, όσον αφορά στα ανοιχτά συνδεδεμένα νομικά δεδομένα, το Semantic Finlex της Φιλανδίας μπορεί να λειτουργήσει ως πρότυπο για τον τρόπο που πρέπει να δομούνται, να δημοσιεύονται και να διασυνδέονται μεταξύ τους τα δεδομένα νομοθεσίας και νομολογίας. Εάν στο μέλλον περισσότερες χώρες εφαρμόσουν το εν λόγω πλήρες και ολοκληρωμένο μοντέλο τότε, με την χρήση των κατάλληλων εργαλείων εξόρυξης και ανάλυσης των δεδομένων, θα έχουμε τη δυνατότητα να εξάγουμε χρήσιμα συμπεράσματα για τον τρόπο άσκησης της πολιτικής σε τομείς (όπως πχ η φορολογία των επιχειρήσεων), για τα ανακλαστικά των κυβερνήσεων στην αντιμετώπιση κοινωνικών φαινομένων (όπως πχ η μετανάστευση) αλλά και για την ταχύτητα απονομής της δικαιοσύνης και κατά πόσο αυτή μπορεί να συσχετιστεί άμεσα ή έμμεσα με άλλους δείκτες (όπως πχ το επενδυτικό ενδιαφέρον).

Όσον δε αφορά στο στατιστικό σετ δεδομένων που επιλέξαμε να αναλύσουμε και απεικονίζει τις αφίξεις και στις δαπάνες των τουριστών κατά το έτος 2017, κάναμε καταρχήν εισαγωγή στην LOD extension του RapidMiner και εν συνεχεία συνδέσαμε το παραπάνω dataset με την DBpedia για να αντλήσουμε δεδομένα σχετιζόμενα με τις χώρες που αναφέρονται στο δείγμα μας. Ακολούθως προσθέσαμε επιπλέον πληροφορίες από την Wikidata, για να διαπιστώσουμε όμως ότι αρκετοί δείκτες είτε είχαν «missing values»

είτε δεν είχαν το σωστό format (πχ αλφαριθμητικό αντί για αριθμητικό). Με την επιλογή των κατάλληλων operators «καθαρίσαμε» τα δεδομένα και προέκυψαν τελικά μόνο 44 μεταβλητές, πολλές εκ των οποίων ήταν κοινές,. Κατά συνέπεια, στο correlation matrix υπήρχε επαναλαμβανόμενη απεικόνιση των ίδιων δεικτών υψηλής και χαμηλής συσχέτισης, γεγονός που περιόρισε την εξαγωγή συμπερασμάτων. Θέλοντας να εμπλουτίσουμε το ως άνω dataset με περισσότερες ενδιαφέρουσες αριθμητικές μεταβλητές, κάναμε περαιτέρω έρευνα σε ανοιχτά δεδομένα και μπήκαμε στην Παγκόσμια Τράπεζα [https://www.worldbank.org/] όπου υπάρχουν περίπου 1600 επικαιροποιημένοι αριθμητικοί δείκτες ανά χώρα. Τα συγκεκριμένα δεδομένα όμως, αν και είναι ελεύθερα προσβάσιμα (open), δεν είναι συνδεδεμένα (linked), ενώ το SPARQL endpoint που υπήρχε ενημερώθηκε τελευταία φορά το 2016 και φαίνεται ανενεργό. Εάν είχαμε λοιπόν την ευχέρεια να συνδέσουμε δείκτες από την Παγκόσμια Τράπεζα όπως πχ ποσά που επενδύονται σε εκπαίδευση, υγεία, υποδομές λιμανιών και αεροδρομίων κ.ο.κ. τότε τα είχαμε διευρύνει περισσότερο το δείγμα μας με πληροφορίες και θα είχαμε τη δυνατότητα, εκτός από correlation analysis, να χρησιμοποιήσουμε και άλλες μεθόδους όπως πχ. clustering.

Το γενικό συμπέρασμα που προέκυψε από την έρευνά μας για την ανάλυση των ανοιχτών συνδεδεμένων δεδομένων, νομικών και στατιστικών, ήταν ότι οι τεχνικές και τα πρότυπα του Σημασιολογικού Ιστού δεν έχουν εφαρμοστεί σωστά, με αποτέλεσμα τα δισεκατομμύρια δεδομένων που παράγονται να μην γίνονται αντικείμενο εκμετάλλευσης και να μην μετουσιώνονται τελικά σε γνώση.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ - ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

Με την παρούσα εργασία μελετήσαμε τα ανοιχτά συνδεδεμένα δεδομένα εστιάζοντας στα ανοιχτά συνδεδεμένα νομικά δεδομένα και τις μέχρι σήμερα πρωτοβουλίες για την ελεύθερη πρόσβαση, λήψη και εκμετάλλευσή τους. Αναφερθήκαμε περαιτέρω στη διαδικασία εξόρυξης και ανάλυσης των δεδομένων με την χρήση του εργαλείου του Rapid Miner και καταγράψαμε διάφορες μεθόδους ανάλυσης για την εξαγωγή γνώσης. Ερευνήσαμε προσεκτικά και εξοικειωθήκαμε με τις οντολογίες της πλατφόρμας δημοσίευσης, λήψης, επαναχρησιμοποίησης και διασύνδεσης νομικών δεδομένων «Semantic Finlex» όπου δημοσιεύεται το Φιλανδικό Δίκαιο και τα σχετιζόμενα έγγραφα (όπως δικαστικές αποφάσεις), σύμφωνα με τις αρχές και τις τεχνικές των Ανοιχτών Συνδεδεμένων Δεδομένων. Προχωρήσαμε στην εκμάθηση της επέκτασης Rapid Miner LOD extension και εκτελέσαμε ερωτήματα SPARQL σε νομικά και στατιστικά σύνολα δεδομένων με τον ορισμό διαφόρων μεταβλητών και τη χρήση των κατάλληλων linkers και operators, ώστε να εκμεταλλευτούμε τα δεδομένα, να προσθέσουμε επιπλέον πληροφορίες, να βρούμε συσχετίσεις και να αναλύσουμε τα ευρήματά μας.

Κατά το αρχικό στάδιο της έρευνας για την αναλυτική των ανοιχτών συνδεδεμένων δεδομένων σε θεωρητικό επίπεδο, δημιουργήθηκε η προσδοκία ότι η εξεύρεση ποιοτικών συνόλων δεδομένων, διαμορφωμένων σύμφωνα με τις αρχές του Σημασιολογικού Ιστού, θα είναι σχετικά εύκολη, δεδομένου ότι σε καθημερινή βάση παράγεται τεράστιος όγκος πληροφοριών. Ωστόσο, κατά την πορεία συγγραφής της εργασίας συναντήσαμε πολλές δυσκολίες στο να εντοπίσουμε σωστά δομημένα, ενημερωμένα και ποιοτικά σετ δεδομένων που θα μπορούσαμε να συσχετίσουμε και να αναλύσουμε ώστε να αποκτήσουμε γνώση. Έχοντας επιλέξει καταρχήν το legal domain για την εξόρυξη και ανάλυση δεδομένων, ερευνήσαμε το www.legislation.gov.uk, για να διαπιστώσουμε ότι είναι δυνατή μόνο η περιήγηση στα δεδομένα μέσω API και δεν παρέχεται η δυνατότητα να κάνουμε λήψη και μεταφόρτωση των δεδομένων ως σύνολο. Μελετήσαμε ακόμη το project EuCases [http://eucases.eu] αλλά δεν το επιλέξαμε διότι τα δεδομένα δεν έχουν επικαιροποιηθεί από την ολοκλήρωση του έργου το 2015 και μετά. Εξερευνήσαμε και το https://eur-lex.europa.eu που όμως δε διαθέτει endpoint για την εκτέλεση ερωτημάτων. Καταλήξαμε λοιπόν στην επιλογή του Semantic Finlex, διότι διαθέτει SPARQL endpoint για την εκτέλεση ερωτημάτων και αναλυτική τεκμηρίωση του RDF μοντέλου για τη

νομοθεσία και τη νομολογία. Μετά την σύνταξη και εκτέλεση των queries με τις κατάλληλες μεταβλητές για τη νομοθεσία και την νομολογία της Φιλανδίας, μπορέσαμε να εξάγουμε συμπεράσματα σχετικά με τους νόμους που έχουν τις περισσότερες versions, την συχνότητα με την οποία τροποποιούνται τα νομοθετήματα και κατά μέσο όρο πόσος χρόνος απαιτείται για την έκδοση δικαστικών αποφάσεων. Όπως αναφέραμε, εάν υπήρχαν τα αντίστοιχα νομικά δεδομένα δημοσιευμένα κατά τις αρχές των Ανοιχτών Συνδεδεμένων Δεδομένων όπως στην Φιλανδία, θα μπορούσαμε να κάνουμε και συγκριτικές αναλύσεις σχετικά με τον ρυθμό μεταβολής των νομοθετημάτων ίδιου αντικειμένου, την ταχύτητα απονομής της δικαιοσύνης, την απόδοση των δικαστικών λειτουργών και να αναλύσουμε τις τυχόν συσχετίσεις αυτών των πληροφοριών με επιπρόσθετες πληροφορίες των χωρών όπως πχ. το ΑΕΠ, την άσκηση πολιτικής, τις καταναλωτικές συνθήκες, τις επενδύσεις κλπ. Στην συνέχεια, θέλοντας να εφαρμόσουμε τις δυνατότητες του εργαλείου RapidMiner LOD extension επιχειρήσαμε αρχικά να εισάγουμε με τη βοήθεια της επέκτασης τα δεδομένα που αντλήσαμε από το Semantic Finlex με τα παραπάνω ερωτήματα. Ωστόσο, υπήρξε σφάλμα που σχετιζόταν με το πρωτόκολλο. Για το λόγο αυτό και επειδή επιμέρους στόχος της εργασίας ήταν να παρουσιάσει πώς θα εμπλουτίσουμε τα ανοιχτά συνδεδεμένα δεδομένα για να παραχθεί γνώση, αλλάξαμε domain και αναζητήσαμε στατιστικά σετ δεδομένων που συνδυάζουν στοιχεία διαφόρων χωρών. Αφού εντοπίσαμε ένα ενδιαφέρον dataset στο statista.com αναφορικά με τις επισκέψεις τουριστών (σε εκατομμύρια) και τα χρήματα που δαπάνησαν (σε δισεκατομμύρια) ανά χώρα κατά το έτος 2017, εισάγαμε τα δεδομένα στο LOD extension του RapidMiner, έχοντας προηγουμένως μελετήσει προσεκτικά τις δυνατότητες των εργαλείων που διαθέτει η επέκταση. Επιλέγοντας τα κατάλληλα εργαλεία της LOD extension προσθέσαμε επιπλέον πληροφορίες και έτσι το αρχικό σετ δεδομένων που περιείχε μόνο δύο τιμές συνδέθηκε πολύ εύκολα με δύο διαφορετικά RDF stores (DBpedia και Wikidata) και προστέθηκαν σε πρώτη φάση 242 επιπλέον μεταβλητές (244 συνολικά) που βρίσκονταν σε άμεση σχέση με την κύρια μεταβλητή του dataset μας που είναι η χώρα. Κατά την ανάλυση της ποιότητας των δεδομένων, διαπιστώσαμε ότι αρκετές από τις 244 μεταβλητές είχαν χαμένες τιμές (missing values) ή δεν είχε οριστεί σωστά ο τύπος των δεδομένων (πχ ακέραιος, πραγματικός κλπ). Μετά από καθάρισμα των μεταβλητών από χαμένες τιμές και μη ορισθέν format, προέκυψαν τελικώς 44 μεταβλητές, αρκετές μάλιστα των οποίων ήταν κοινές σε DBpedia και Wikidata. Εν συνεχεία προχωρήσαμε σε ανάλυση συσχέτισης των δεδομένων (correlation) και εξάγαμε συμπεράσματα. Θα πρέπει να σημειώσουμε ότι αφού διαπιστώσαμε ότι τα αποτελέσματα ήταν ελλιπή και είχαν πολλές επαναλαμβανόμενες

τιμές, μπήκαμε στο https://www.worldbank.org/ όπου υπάρχουν περίπου 1600 επικαιροποιημένοι αριθμητικοί δείκτες ανά χώρα. Τα δεδομένα στο https://www.worldbank.org/ είναι μεν ελευθέρα προσβάσιμα (open) αλλά δεν είναι συνδεδεμένα (linked), ενώ το SPARQL endpoint που υπήρχε ενημερώθηκε τελευταία φορά το 2016 και φαίνεται ανενεργό. Εάν λοιπόν μπορούσαμε να συνδέσουμε ενδιαφέροντες αριθμητικούς δείκτες του https://www.worldbank.org/ όπως πχ ποσά που επενδύονται σε εκπαίδευση, υγεία, υποδομές λιμανιών και αεροδρομίων κ.ο.κ. τότε τα είχαμε εμπλουτίσει ακόμη περισσότερο το δείγμα μας με πληροφορίες και θα είχαμε τη δυνατότητα, εκτός από correlation analysis, να κάνουμε και clustering.

Διαπιστώθηκε λοιπόν κατά την εκπόνηση της παρούσας εργασίας ότι σε πολλές περιπτώσεις οι δημιουργοί και εκδότες των διαφόρων κατηγοριών ανοιχτών δεδομένων δεν είχαν εφαρμόσει τα πρότυπα μορφοποίησης και τις οδηγίες δημοσίευσης του W3C, με αποτέλεσμα είτε να μην λειτουργεί η σύνδεση με τα endpoints, είτε οι ίδιες μεταβλητές να λαμβάνουν διαφορετικό format (πχ string και integer για την ημερομηνία), είτε ακόμη να υπάρχουν missing values σε τιμές μεταβλητών για την ίδια οντότητα σε δύο διαφορετικές πηγές πχ Dbpedia και Wikidata. Διαπιστώσαμε ακόμη την έλλειψη ποιοτικών datasets ανοιχτών δεδομένων και την απουσία διαλειτουργικότητας μεταξύ των πηγών ανοιχτών δεδομένων που συχνά εμποδίζει την επαναχρησιμοποίησή τους και κυρίως την εκμετάλλευσή τους. Παρόλα αυτά υπάρχουν τομείς όπου τα πρότυπα και οι τεχνικές των ανοιχτών συνδεδεμένων δεδομένων έχουν εφαρμοστεί σωστά, όπως τα γεωγραφικά δεδομένα και τα ιατρικά δεδομένα.

Αξιολογώντας τα παραπάνω, αντιλαμβανόμαστε την σωστή προσέγγιση της πλατφόρμας LDF.fi της Φινλανδίας (όπου δημοσιεύονται τα νομικά συνδεδεμένα δεδομένα του Semantic Finlex), η οποία υιοθετεί την άποψη ότι για να είναι δυνατή η ουσιαστική χρήση των συνόλων δεδομένων, εκτός των 5 αστέρων του Tim Berners-Lee, χρειάζεται επιπλέον: i) παροχή των δεδομένων με ένα σαφές σχήμα και τεκμηρίωση, έτσι ώστε οι άνθρωποι να μπορούν εύκολα να κατανοούν και να επαναχρησιμοποιούν τα δεδομένα και τα μεταδεδομένα και ii) επαλήθευση των δεδομένων και δήλωση της προέλευσής τους έτσι ώστε οι άνθρωποι να μπορούν να εμπιστεύονται την ποιότητά τους (μοντέλο 7 αστέρων). Πράγματι, εάν υπήρχε μεγαλύτερη υποστήριξη των καταναλωτών, καλύτερη πρόσβαση στα δεδομένα και πληρέστερη κατανόηση της ποιότητας και προέλευσής, η επιστήμη της εξόρυξης και ανάλυσης των ανοιχτών συνδεδεμένων δεδομένων θα

διαδραμάτιζε καθοριστικό ρόλο στην λήψη στρατηγικών αποφάσεων και στην απόκτηση γνώσης.

Με την ανάπτυξη του «διαδικτύου των πραγμάτων» (Internet of Things – IoT), οδηγούμαστε στη συλλογή και παραγωγή ακόμη περισσότερων δεδομένων από τις συνδεδεμένες έξυπνες συσκευές, τα οποία θα προστεθούν στα δισεκατομμύρια bytes δεδομένων που παράγονται καθημερινά, κάνοντας ακόμη πιο ξεκάθαρη την ανάγκη για αξιοποίηση και ανάλυση των πληροφοριών.

Μια πρόταση για μελλοντική έρευνα προς το σκοπό επίλυσης των προεκτεθέντων προβλημάτων θα ήταν η εξής: καθότι τα ανοιχτά δεδομένα είναι διαθέσιμα από διάφορες ετερογενείς πλατφόρμες όπως το LOD cloud, η Dbpedia, η Wikidata, οι παραπάνω δυσχέρειες θα μπορούσαν να επιλυθούν με την τεχνική της ενοποίησης, δηλαδή με το να δημιουργηθούν γέφυρες για την ενσωμάτωση αυτών των πλατφορμών και τη δημιουργία πιο εκτεταμένων ιστών δεδομένων. Μια ακόμη ιδέα θα ήταν η δημιουργία ενός ενιαίου RDF store ανά domain όπου θα εφαρμοζόταν αυστηροί κανόνες και συχνή εποπτεία ως προς την πληρότητα, ποιότητα και την ενημέρωσή τους και θα υπήρχε πρόσβαση με βάση μία οντολογία ανά domain που θα μπορούσε να επεκταθεί με συμφωνία όλων των εμπλεκομένων φορέων.

Με τον Σημασιολογικό Ιστό και τα ανοιχτά συνδεδεμένα δεδομένα (Web 3.0) να ανοίγουν πλέον το δρόμο για το λεγόμενο Symbiotic Web (Web 4.0) αναμένεται να αναπτυχθεί λογισμικό που θα επεξεργάζεται τα δεδομένα και θα κάνει την κατάλληλη επιλογή ανάλογα με τις ανάγκες και το προφίλ του εκάστοτε χρήστη. Υπό αυτές τις συνθήκες, η εκμετάλλευση και ανάλυση των ανοιχτών συνδεδεμένων δεδομένων αναμένεται να διαδραματίσει βασικό ρόλο στον τρόπο απόκτησης γνώσης και λήψης αποφάσεων σε όλους τους τομείς της ζωής.

ΠΗΓΕΣ – ΒΙΒΛΙΟΓΡΑΦΙΑ

Auer, S., et al. (2009). Triplify – Light-Weight Linked Data Publication from Relational Databases. Proceedings of the 18th World Wide Web Conference (WWW2009).

Begum, S.H, (2013). Data Mining Tools and Trends – An Overview, International Journal of Emerging Research in Management & Technology, ISSN: 2278-9359, p.6-12.

Bekker, A., (2017): 4 types of data analytics to improve decision-making, ανακτήθηκε από https://www.scnsoft.com/blog/4-types-of-data-analytics.

Berners-Lee, T. (2009). Linked Data, ανακτήθηκε από https://www.w3.org/DesignIssues/LinkedData.html.

Berrueta, D., Phipps, J., (2008). Best Practice Recipes for Publishing RDF Vocabularies-W3C Working Group Note, ανακτήθηκε από https://www.w3.org/TR/swbp-vocab-pub/.

Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked Data – The story so far, International Journal on Semantic Web and Information Systems, ανακτήθηκε από http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf .

Bizer, C., Cyganiak, R., Health, T. (2007). How to publish Linked Data on the Web, ανακτήθηκε από http://wifo5-03.informatik.uni-mannheim.de/bizer/pub /LinkedDataTutorial/.

Boella, G., et al., (2015). Linking Legal Open Data: Breaking the Accessibility and language barrier in European Legislation and Case Law, International Conference on Artificial Intelligence and Law 2015, pp. 171-175, ανακτήθηκε από https://www.researchgate.net/publication/300403178_Linking_legal_open_data.

Boer, A., Hoekstra, R., Winkels, R., Van Engers, T., Willaert, F. (2002). MET Alex: Legislation in XML, JURIX: The Fifteenth Annual Conference, London.

Carroll, J., Bizer, C., Hayes, P., Stickler, P. (2005). Named graphs. Journal of Web Semantics, 3(4):247-267.

Chalkidis, I., Nikolaou, C., Soursos, P., Koubarakis, M., (2017). Modeling and Querying Greek Legislation using Semantic Web Technologies, Springer, 2017, ανακτήθηκε από https://pdfs.semanticscholar.org/11ec/ebda5bbcba4dee9292d481059932d8a57460.pdf.

Coetzee, P., Heath, T., Motta, E. (2008). SparqPlug. Proceedings of the 1st Workshop on Linked Data on the Web (LDOW2008).

Council of European Union, (2012). Council Conclusions Inviting the Introduction of the European Legislation Identifier (ELI). Technical Report Official Journal Issue C

325, ανακτήθηκε από https://eur-lex.europa.eu/legal-content/EN/ ALL/?uri= CELEX %3A52012XG1026%2801%29.

Domo.com (2018), 6th Annual Data never sleeps infographic, ανακτήθηκε από https://www.domo.com/news/press/domo-releases-sixth-annual-data-never-sleeps-infographic.

EuCases Project, (2014). Report on LT2XML Conversion Tools, ανακτήθηκε από http://eucases.eu/fileadmin/eucases/documents/eucases_d2.3_lt2xmltools_report.pdf.

EuCases Project, (2015). EUCases exploitation planning report, ανακτήθηκε από http://leadership2017.eu/fileadmin/EUCases/documents/Deliverables/EUCases_D6.7_Exploitation_Report_final.pdf

European Commission (2011). Open Data: An engine for innovation, growth and transparent governance, Communication from the Commission, COM, ανακτήθηκε από http://www.europarl.europa.eu/RegData/docs_autres_institutions /commission_ europeenne/com/2011/0882/COM_COM(2011)0882_EN.pdf.

Francesconi, E., (2006). The "Norme In Rete" – Project: Standards and Tools for Italian Legislation. International Journal of Legal Information, τεύχος 34, pp. 358-3756, ανακτήθηκε από https://scholarship.law.cornell.edu/ cgi/viewcontent.cgi? article =1061&context=ijli.

Francesconi, E., (2018). On the Future of Legal Publishing Services in the Semantic Web, Future Internet – Open Access Journal, ανακτήθηκε από www.mdpi.com/1999-5903/10/6/48/pdf.

Frosterus, M., Tuominen, J., Hyvonen, E., (2014). Facilitating Re-use of Legal Data in applications – Finnish Law as a Linked Open Data Service, JURIX 2014, pp. 115-124, ανακτήθηκε από http://docplayer.net/52587928-Facilitating-re-use-of-legal-data-in-applications-finnish-law-as-a-linked-open-data-service.html

Hartig, O., (2009). Provenance Information in the Web of Data. Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009).

Heath, T, Bizer, C. (2011). Linked data: Evolving the web into a global data space. Synthesis Lectures on the Semantic Web: Theory and Technology 1(1), 1–136.

Hirudkar, A., Sherekar, S., (2013). Comparative Analysis of Data Mining Tools and Techniques for evaluating of database systems, International Journal Of Computer Science And Applications Vol. 6, No.2, ISSN: 0974-1011 (Open Access).

Hofmann, M., Klinkenberg, R., (2013). RapidMiner – Data mining use cases and business analytics applications, CRC Press, Taylor & Francis Group.

Janssen, F., Fallahi, F., Nößner, I., Paulheim, H. (2012). Towards Rule Learning Approaches to Instance-based Ontology Matching. In: Proceedings of the First

International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data; 13-18. RWTH, Aachen.

KDNuggets, (2018). 19th annual KDnuggets Software Poll, ανακτήθηκε από https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

Klyne, G., Caroll, J., (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax – W3C recommendation, ανακτήθηκε από https://www.w3.org/TR/rdf-concepts/.

Lausch, A., Schmidt, A., Tischendorf, L. (2015). Data mining and linked open data – New perspectives for data analysis in environmental research, Ecological Modelling Elsevier, vol. 295(C), pages 5-17.

Oksanen, A., et al., (2017). Law and Justice as a Linked Open Data Service, ανακτήθηκε από https://seco.cs.aalto.fi/publications/submitted/law-justice-linked.pdf

Oksanen, A., et al., (2018). Semantic Finlex: Finnish Legislation and Case Law as Linked Open Data Service, ανακτήθηκε από https://seco.cs.aalto.fi/publications/2018/oksanen-et-al-lvi-2018.pdf

Paulheim, H., (2015). Linked Open Data enhanced Knowledge Discovery. Introducing the RapidMiner Linked Open Data Extension ανακτήθηκε από https://www.slideshare.net/heikopaulheim/linked-open- data- enhanced-knowledge-discovery

Paulheim, H., Ristoski, Mitichkin, E., Bizer, C., (2014): Data Mining with Background Knowledge from the Web. In: RapidMiner World, 2014.

Resource Description Framework (RDF), (n.d.). Concepts and Abstract Syntax, ανακτήθηκε από https://www.w3.org/TR/2004/REC-rdf-concepts-20040210.

Ristoski, P. (2015). Towards Linked Open Data enabled Data Mining. In The Semantic Web. Latest Advances and New Domains, pages 772–782. Springer.

Ristoski, P., Bizer, C., Paulheim, H., (2015). Mining the Web of Linked Data with RapidMiner, Journal of Web Semantics, Volume 35, Issue P3, pages 142-151.

Sauermann, L., Cyganiak, R., (2008). Cool URIs for the Semantic Web-W3C Interest Group Note, ανακτήθηκε από http://www.w3.org/TR/cooluris/.

Schmachtenberg, M., Bizer, C., Paulheim, H., (2014). Adoption of the Linked Data Best practices in Different Topical Domains, ISWC 2014, LNCS 8796, Springer international Publishing Switzerland, ανακτήθηκε από https://pdfs.semanticscholar.org/aeaa/0dd7d2bd396e458d0f442d4146e9c3c4d0bc.pdf.

Spinosa, P., Francesconi, E., Lupo, C., (2018). A Uniform Resource Name (URN) for Sources of Law (LEX), Technical Report IETF, ανακτήθηκε από https://datatracker.ietf.org/doc/draft-spinosa-urn-lex/?include_text=1.

Stroetmann, Veli N. (2013). Linking Legal Open Data- An Innovative EU Approach, EuCases Project, ανακτήθηκε από https://www.ajbd.de/wp-content/uploads/2013/09/AjBD_2013_Stroetmann.pdf

van Assem, M, (2010), Converting and Integrating Vocabularies for the Semantic Web, SIKS Dissertation Series No. 2010-40, Volume 295, 10 January 2015, Pages 5-17.

W3C organization, (2017). SweoIG/ TaskForces/ Community Projects / LinkingOpenData, ανακτήθηκε από https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/ LinkingOpenData#Project_Description.

W3C organization, (n.d). Cool URIs don't change, ανακτήθηκε από https://www.w3.org/Provider/Style/URI.

Wikipedia, (n.d), SPARQL, ανακτήθηκε από https://en.wikipedia.org/wiki/SPARQL

Wikipedia, (n.d.). *Linked Data*, ανακτήθηκε από https://el.wikipedia.org/wiki/Linked_Data.

Witten, I., Eibe, F, Hall, M. (2011). Practical Machine Learning Tools and Techniques, 3rd edition, Morgan Kaufmann Publishers.