



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

**Οι Συζητήσεις Του Ελληνικού Κοινοβουλίου Ως Ανοιχτά
Διασυνδεδεμένα Δεδομένα**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΠΑΠΑΝΙΚΟΛΑΟΥ ΙΩΑΝΝΗ

Επιβλέπων : Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Μάριος Κόνιαρης
Ε.ΔΙ.Π. ΕΜΠ

Αθήνα, Φεβρουάριος 2024

Η σελίδα αυτή είναι σκόπιμα λευκή.



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

**Οι Συζητήσεις Του Ελληνικού Κοινοβουλίου Ως Ανοιχτά
Διασυνδεδεμένα Δεδομένα**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΙΩΑΝΝΗ ΠΑΠΑΝΙΚΟΛΑΟΥ

Επιβλέπων : Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Μάριος Κόνιαρης
Ε.ΔΙ.Π. ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 23^η Φεβρουαρίου 2024.

(Υπογραφή)

.....
Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.

(Υπογραφή)

.....
Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

(Υπογραφή)

.....
Ευγενία Τζαννίνη
Επίκουρη Καθηγήτρια Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2024

(Υπογραφή)

.....

ΙΩΑΝΝΗΣ ΠΑΠΑΝΙΚΟΛΑΟΥ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © **ΙΩΑΝΝΗΣ, ΠΑΠΑΝΙΚΟΛΑΟΥ, 2024.**

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Το Ελληνικό Κοινοβούλιο παράγει μεγάλες ποσότητες πολύτιμων κειμενικών δεδομένων κατά τη διάρκεια των νομοθετικών συνόδων, με την διάχυση τους να γίνεται σε μορφές όπως PDF, DOC(X) ή TXT. Οι υπάρχουσες μέθοδοι πρόσβασης και ανάλυσης αυτών των δεδομένων αντιμετωπίζουν προκλήσεις λόγω της μη δομημένης φύσης τους. Ως απάντηση, η παρούσα εργασία παρουσιάζει μια λύση που αξιοποιεί τεχνολογίες σημασιολογικού ιστού και συνδεδεμένων δεδομένων για τη δημιουργία μιας δομημένης αναπαράστασης που διευκολύνει την εξειδικευμένη αναζήτηση και ανάλυση.

Πιο αναλυτικά, αναπτύσσουμε ένα ολοκληρωμένο σύστημα συνδεδεμένων δεδομένων προσαρμοσμένο στις ιδιαιτερότητες των κοινοβουλευτικών συζητήσεων. Η προσέγγιση αυτή αποσκοπεί στην ενίσχυση της προσβασιμότητας τους αξιοποιώντας κατάλληλα την πληροφορία από το αδόμητο υλικό. Μετατρέπουμε τα αδόμητα αρχεία σε δομημένες μορφές και επεκτείνουμε τη σημασιολογική αναπαράστασή τους δημιουργώντας τριπλέτες RDF. Για να εμπλουτίσουμε τα σημασιολογικά μας δεδομένα δημιουργούμε διασυνδέσεις με το Wikidata, ένα δημοφιλές σύνολο γνώσης. Στην συνέχεια, επεξεργαζόμαστε και αναλύουμε το σύνολο των σημασιολογικών αρχείων με την αξιοποίηση συστήματος διαχείρισης τέτοιων αρχείων.

Οι πρωταρχικοί στόχοι αυτής της εργασίας περιλαμβάνουν τη διευκόλυνση της ευκολότερης ανάλυσης, την εξαγωγή κρίσιμων πληροφοριών και την δυνατότητα εκτέλεσης σημασιολογικών ερωτημάτων. Μετατρέποντας τα ακατέργαστα αρχεία κειμένου σε δομημένες μορφές, το σύστημά μας συμβάλλει στην πληρέστερη κατανόηση των νομοθετικών δραστηριοτήτων στο Ελληνικό Κοινοβούλιο. Η εργασία αυτή δεν αντιμετωπίζει μόνο τις τεχνικές προκλήσεις που σχετίζονται με τη διαλειτουργικότητα των δεδομένων, αλλά υπογραμμίζει επίσης τη σημασία των τεχνολογιών του σημασιολογικού ιστού για την προώθηση της διαφάνειας και της αποτελεσματικότητας στις κοινοβουλευτικές διαδικασίες.

Λέξεις Κλειδιά: Σημασιολογικός Ιστός, Συνδεδεμένα Δεδομένα, Ελληνικό Κοινοβούλιο

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

The Greek Parliament produces large amounts of valuable textual data during legislative sessions, which are disseminated in formats such as PDF, DOC(X) or TXT. Existing methods for accessing and analyzing this data face challenges due to its unstructured nature. In response, this thesis presents a solution that utilizes semantic web and connected data technologies to create a structured representation that facilitates specialized search and analysis.

More specifically, we develop an integrated linked data system tailored to the specificities of parliamentary debates. This approach aims to enhance their accessibility by appropriately exploiting the information from the unstructured material. We convert the unstructured files into structured formats and extend their semantic representation by creating RDF triples. To enrich our semantic data we create links to Wikidata, a popular knowledge set. Then, we process and analyze the set of semantic files by utilizing a system for managing such files.

The primary goals of this thesis include facilitating easier analysis, extracting critical information and enabling the execution of semantic queries. By converting raw text files into structured formats, our system contributes to a more complete understanding of legislative activities in the Greek Parliament. This diploma thesis not only addresses the technical challenges related to data interoperability, but also highlights the importance of semantic web technologies for promoting transparency and efficiency in parliamentary processes.

Keywords: Semantic Web, Linked Data, Greek Parliament

Η σελίδα αυτή είναι σκόπιμα λευκή.

Ευχαριστίες

Ξεκινώντας, θα ήθελα να εκφράζω τη βαθύτατη ευγνωμοσύνη μου στον κύριο Μάριο Κόνιαρη. Οι εβδομαδιαίες συναντήσεις πρόσφεραν πολύτιμη καθοδήγηση. Οι γνώσεις και η επιμονή του με βοήθησαν να βελτιώσω τις ιδέες μου και να ολοκληρώσω την διπλωματική μου εργασία. Η ευγνωμοσύνη μου δεν γίνεται να μην επεκταθεί και στον κύριο Παναγιώτη Τσανάκα, που με εμπιστεύτηκε να φέρω εις πέρας αυτό το απαιτητικό θέμα που διαπραγματεύεται σε αυτήν την εργασία.

Στη συνέχεια, θα ήθελα να ευχαριστήσω τους γονείς μου, Θωμά και Αγγελική, καθώς και τον αδελφό μου Κωνσταντίνο, οι οποίοι στάθηκαν δίπλα μου, δίνοντας μου ατελείωτη υποστήριξη, αγάπη και κατανόηση κατά τη διάρκεια όλων των ακαδημαϊκών μου χρόνων.

Τέλος, θέλω να ζητήσω ένα ευχαριστώ και από τους φίλους μου, των οποίων η παρέα τους συνέβαλε στη δημιουργία μιας ευχάριστης ατμόσφαιρας που με βοήθησε να ξεπεράσω ακόμη και τις πιο δύσκολες και πιεστικές ημέρες.

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας Περιεχομένων

Περίληψη	5
Abstract	7
Ευχαριστίες	9
Πίνακας Περιεχομένων	11
Πίνακας Εικόνων	13
Κατάλογος Πινάκων	15
1 Εισαγωγή.....	16
1.1 Περιγραφή Προβλήματος	16
1.2 Αντικείμενο διπλωματικής.....	16
1.2.1 Συνεισφορά	17
1.3 Σχετικά Παραδείγματα.....	17
1.4 Οργάνωση κειμένου	18
2 Θεωρητικό υπόβαθρο.....	20
2.1 Σημασιολογικός Ιστός – Semantic Web	20
2.2 Ανοιχτά Διασυνδεδεμένα Δεδομένα	22
2.3 Τεχνολογίες Υποστήριξης Δεδομένων	24
2.3.1 Extensible Markup Language (XML)	25
2.3.2 Resource Description Framework (RDF).....	26
2.3.3 Λεξιλόγια Σημασιολογικού Ιστού.....	27
2.3.4 SPARQL.....	29
2.4 Ανοιχτά Κυβερνητικά Δεδομένα – LegalDocML	30
2.5 TEI ParlaMint	32
2.6 Θεματική Μοντελοποίηση (Topic Modelling) - LDA	32
3 Σύστημα Διαχείρισης Πρακτικών	34
3.1 Αρχιτεκτονική Συστήματος	34

3.2	Δεδομένα Συστήματος	36
3.2.1	Βάση Δεδομένων	36
3.2.2	Δομή Πρακτικών Βουλής.....	37
3.3	Διαχείριση και Μετατροπή Δεδομένων	40
3.3.1	Εξαγωγή και Ανάλυση Κειμένων.....	41
3.3.1.1	Γλώσσα Περιγραφής Κοινοβουλευτικών Αρχείων.....	41
3.3.2	Μετατροπή Αρχείων Κειμένου σε XML αρχεία.....	43
3.3.2.1	Διασύνδεση με Wikidata	45
3.3.3	Μετατροπή Αρχείων XML σε RDF αρχεία	46
3.3.3.1	RDF διασύνδεση με Wikidata.....	49
3.3.3.2	RDF με κυβερνητικά δεδομένα από επίσημες πηγές	49
3.3.4	Μετατροπή Αρχείων XML LegalDocML σε TEI ParlaMint.....	50
3.4	Πρόσβαση και Αναζήτηση δεδομένων RDF - Apache Jena Fuseki	51
3.5	Σχολιασμός Διαδικασίας Μετατροπής.....	52
4	Μελέτη Περιπτώσεων και Στατιστικά Δεδομένα	54
4.1	Περίπτωση μελέτης RDF	54
4.2	Περίπτωση μελέτης στα δομημένα αρχεία – LDA Topic Modeling	60
4.3	Στατιστικά Δεδομένα	65
5	Επίλογος	68
5.1	Σύνοψη και συμπεράσματα.....	68
5.2	Μελλοντικές επεκτάσεις	69
6	Βιβλιογραφία	71

Πίνακας Εικόνων

Εικόνα 2.1 – Η εξέλιξη του Διαδικτύου	21
Εικόνα 2.2 – The Linked Open Data Cloud	23
Εικόνα 2.3 – Χρονική ανάπτυξη των συνδεδεμένων ανοικτών δεδομένων.....	24
Εικόνα 2.4 – Βασική σύνταξη RDF	26
Εικόνα 2.5 – RDF(S) παράδειγμα	28
Εικόνα 2.6 – Ενδεικτικό απλό ερώτημα σε γλώσσα Sparql.....	30
Εικόνα 3.1 – Διάγραμμα Αρχιτεκτονικής Συστήματος.....	35
Εικόνα 3.2 – Συνολικό pipeline Συστήματος.....	36
Εικόνα 3.3 – Pipeline διαδικασίας μετασχηματισμού	36
Εικόνα 3.4 – Επίσημος ιστότοπος Ελληνικής Κυβέρνησης με Συνεδριάσεις Ολομέλειας	37
Εικόνα 3.5 – Ενδεικτικό απόσπασμα πρακτικών – «ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ» (08-06-2018)	38
Εικόνα 3.6 – Ενδεικτικό απόσπασμα πρακτικών – «Θεμάτων» (08-06-2018)	38
Εικόνα 3.7 – Ενδεικτικό απόσπασμα πρακτικών – «Προεδρεύοντες» και «Ομιλητές» (08-06-2018)	39
Εικόνα 3.8 – Ενδεικτικό απόσπασμα πρακτικών – Εισαγωγικός Πρόλογος (08-06-2018)	39
Εικόνα 3.9 – Ενδεικτικό απόσπασμα πρακτικών – Μέρος Διαλόγων (08-06-2018)	40
Εικόνα 3.10 – Ενδεικτικό δέντρο που προκύπτει από την γραμματική	42
Εικόνα 3.11 – Απόσπασμα από το συντακτικό της γραμματικής	42
Εικόνα 3.12 – Απόσπασμα από τα μεταδεδομένα ενός αρχείου LegalDocML	43
Εικόνα 3.13– Απόσπασμα από κύριο μέρος (debateBody) ενός αρχείου LegalDocML.....	44
Εικόνα 3.14 – Διάγραμμα για το στοιχείο “debate”	44

Εικόνα 3.15 – Σημαντικοί όροι στην Wikidata	45
Εικόνα 3.16(α)-(δ) – Τριπλέτες RDF όπως φαίνονται σε ένα rdf/xml αρχείο	47
Εικόνα 3.17 – Σημασιολογικό μοντέλο για το σύνολο δεδομένων των συζητήσεων	48
Εικόνα 3.18 – Απόσπασμα από κύριο μέρος ενός αρχείου ParlaMint	50
Εικόνα 3.19 – Περιβάλλον υποβολής ερωτημάτων SPARQL	51
Εικόνα 3.20 – Δείγμα απάντησης ερωτημάτων SPARQL σε μορφή πίνακα	52
Εικόνα 4.1 – Παράδειγμα χρήσης – ερώτημα 1 με «Αλέξιος/Αλέξης Τσίπρα».....	55
Εικόνα 4.2 – Παράδειγμα χρήσης – αποτέλεσμα στο ερώτημα 1 με «Αλέξιος/Αλέξης Τσίπρα»	55
Εικόνα 4.3 – Παράδειγμα χρήσης – ερώτημα 2 με βουλευτές Νέας Δημοκρατίας	56
Εικόνα 4.4 – Παράδειγμα χρήσης – αποτέλεσμα στο ερώτημα 2 με βουλευτές της Νέας Δημοκρατίας.....	56
Εικόνα 4.5 – Ερωτήματα ομιλιών και ομιλητών ανά φύλο.....	57
Εικόνα 4.6 – Διαγράμματα ομιλητών και ομιλιών ανά φύλο στις συζητήσεις του Ελληνικού Κοινοβουλίου	58
Εικόνα 4.7 – Διάγραμμα μέσου χρόνου ομιλίας για όλες τις ημερομηνίες.....	59
Εικόνα 4.8 –Κύρια οθόνη Topic Modeling αποτελέσματος με LDA για έτος 2016.....	61
Εικόνα 4.9 – Οθόνη Topic Modeling αποτελέσματος με LDA για έτος 2016 με επιλεγμένο το θέμα 1	61

Κατάλογος Πινάκων

ΠΙΝΑΚΑΣ 4.1 – Αριθμός ομιλητών και ομιλιών ανά φύλο στις συζητήσεις του Ελληνικού Κοινοβουλίου	57
ΠΙΝΑΚΑΣ 4.2 – Αριθμός ομιλητών-βουλευτών ανά κόμμα και φύλο στις συζητήσεις του Ελληνικού Κοινοβουλίου	58
ΠΙΝΑΚΑΣ 4.3 – Πρώτοι δέκα σημαντικοί όροι για κάθε έτος (από μοντέλο LDA).....	63
ΠΙΝΑΚΑΣ 4.4 – Ετήσια στατιστικά στοιχεία μετατροπής αρχικών αρχείων σε XML	65
ΠΙΝΑΚΑΣ 4.5 – Στατιστικά μετατροπής XML ανά κοινοβουλευτική περίοδο	66

1

Εισαγωγή

1.1 Περιγραφή Προβλήματος

Η παρούσα διπλωματική εργασία επικεντρώνεται στην αξιοποίηση των αρχών των συνδεδεμένων δεδομένων και των τεχνολογιών του σημασιολογικού ιστού για τη βελτίωση της προσβασιμότητας και της ανάλυσης των συζητήσεων του Ελληνικού Κοινοβουλίου. Το Ελληνικό Κοινοβούλιο, ως το ανώτατο νομοθετικό όργανο στην Ελλάδα, παράγει έναν τεράστιο όγκο πολύτιμων κειμενικών δεδομένων κατά τη διάρκεια των συζητήσεών του. Ωστόσο, οι υπάρχουσες μέθοδοι πρόσβασης και αξιοποίησης αυτών των δεδομένων παρουσιάζουν αρκετές προκλήσεις και περιορισμούς.

Ο ιστότοπος του Ελληνικού Κοινοβουλίου χρησιμεύει ως η κύρια πηγή πληροφοριών σχετικά με τις κοινοβουλευτικές διαδικασίες. Παρέχει πρόσβαση σε απομαγνητοφωνημένα κείμενα και αρχεία των συζητήσεων, επιτρέποντας στους πολίτες, τους ερευνητές και τους πρωταγωνιστές της πολιτικής σκηνής να ενημερώνονται για τις νομοθετικές δραστηριότητες. Ωστόσο, η τρέχουσα δομή του ιστότοπου υποστηρίζει κυρίως την παραδοσιακή περιήγηση και αναζήτηση με βάση την ημερομηνία, χωρίς προηγμένα χαρακτηριστικά για την εξερεύνηση δεδομένων, τη διασύνδεση και τη σημασιολογική αναζήτηση.

Ένα από τα κύρια προβλήματα που αντιμετωπίζει ο υφιστάμενος δικτυακός τόπος είναι η διάθεση του υλικού μέσα από αδόμητα κείμενα. Οι συζητήσεις διατίθενται στους χρήστες σε μορφή pdf, doc(x) ή txt, γεγονός που καθιστά δύσκολη την εξαγωγή πληροφοριών ή την ανάλυση σε λεπτομερές επίπεδο. Αυτός ο περιορισμός εμποδίζει την ολοκληρωμένη έρευνα και την ικανότητα να γίνουν συνδέσεις μεταξύ διαφορετικών συζητήσεων, θεμάτων ή συμμετεχόντων/ομιλητών.

1.2 Αντικείμενο διπλωματικής

Αντικείμενο της παρούσας διπλωματικής είναι η μετατροπή των νομοθετικών συζητήσεων σε ανοικτά διασυνδεδεμένα δεδομένα, με κύρια έμφαση στη δημιουργία μιας

δομημένης αναπαράστασής τους. Τα δομημένα νομοθετικά αρχεία αποτελούν την βάση ώστε να διευκολυνθεί η πρόσβαση στις συζητήσεις του Κοινοβουλίου, καθιστώντας πιο κατανοητό το νομοθετικό περιεχόμενο.

Μέσω αυτής της διπλωματικής αποσκοπούμε στην δημιουργία ενός συστήματος που μετατρέπει τις συζητήσεις του Ελληνικού Κοινοβουλίου σε συνδεδεμένα δεδομένα χρησιμοποιώντας τεχνολογίες του σημασιολογικού ιστού. Τα μη δομημένα κείμενα τα αναλύουμε με έναν parser που κατασκευάσαμε για αυτόν τον σκοπό και τα μετατρέπουμε σε δομημένα έγγραφα XML. Στη συνέχεια, μετατρέπουμε τα έγγραφα XML σε μορφή σχήματος RDFS (Resource Description Framework Schema), ένας τρόπος δόμησης δεδομένων που επιτρέπει την αναπαράσταση εννοιών, σχέσεων και μεταδεδομένων που σχετίζονται με τις συζητήσεις.

Σκοπός είναι να δημιουργηθεί μια αναπαράσταση συνδεδεμένων δεδομένων των συζητήσεων του Ελληνικού Κοινοβουλίου, η οποία επιτρέπει τη σημασιολογική αναζήτηση. Αυτή η προσέγγιση διευκολύνει τις προηγμένες αναζητήσεις, τα σύνθετα ερωτήματα και την εξαγωγή ουσιαστικών πληροφοριών από το σύνολο δεδομένων. Επιπλέον, δημιουργούμε ένα σχήμα RDFS ειδικά για τον τομέα των κοινοβουλευτικών συζητήσεων της Ελλάδας, αποτυπώνοντας την απαραίτητη γνώση για ολοκληρωμένη ανάλυση.

1.2.1 Συνεισφορά

Στη παρούσα εργασία, αναλύουμε, επεξεργαζόμαστε και έπειτα μετατρέπουμε τα πρακτικά του Ελληνικού Κοινοβουλίου σε δομημένες μορφές. Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

- Αναλύουμε τα αρχεία των συζητήσεων του Ελληνικού Κοινοβουλίου. Με την κατάλληλη επεξεργασία δημιουργούμε δομημένα αρχεία της μορφής LegalDocML και ParlaMint.
- Επεκτείνουμε την σημασιολογική αναπαράσταση των αρχείων με την δόμησή τους σε αρχεία RDF μορφής. Αυτός ο μετασχηματισμός επιτρέπει προηγμένα σημασιολογικά ερωτήματα και αναλύσεις στο πλαίσιο του συστήματος.
- Παρέχουμε πρόσβαση στην δομημένη σημασιολογική αναπαράσταση των κοινοβουλευτικών πρακτικών, για αναλυτική έρευνα και μελέτη.
- Παρουσιάζουμε τα αποτελέσματα με μελέτη περιπτώσεων στα δομημένα αρχεία και στατιστικά δεδομένα αναφορικά με τα πρακτικά του Ελληνικού Κοινοβουλίου.

1.3 Σχετικά Παραδείγματα

Προκειμένου να κατανοήσουμε την εφαρμογή των συνδεδεμένων δεδομένων στις κοινοβουλευτικές διαδικασίες αξίζει να εξετάσουμε σχετικά διεθνή παραδείγματα. Τα Πρακτικά του Ευρωπαϊκού Κοινοβουλίου και τα Πρακτικά του Κοινοβουλίου του Καναδά είναι δύο γνωστά παραδείγματα που εξετάζονται στην παρούσα ενότητα.

Τα Πρακτικά του Ευρωπαϊκού Κοινοβουλίου¹ είναι ένα έργο που χρησιμοποιεί τις αρχές των συνδεδεμένων δεδομένων για να παρέχει πρόσβαση στις κοινοβουλευτικές συζητήσεις και τις σχετικές πληροφορίες. Οι χρήστες μπορούν να περιηγηθούν και να εξετάσουν συζητήσεις, ομιλίες, αρχεία ψηφοφορίας και άλλες νομοθετικές δραστηριότητες στον ιστότοπο, ο οποίος

¹ <https://linkedpolitics.ops.few.vu.nl/web/html/home.html>

παρέχει ένα πλήρες αρχείο των κοινοβουλευτικών διαδικασιών. Το σύνολο δεδομένων περιλαμβάνει κάθε συζήτηση στην ολομέλεια του Ευρωπαϊκού Κοινοβουλίου (ΕΚ) από τον Ιούλιο του 1999 έως τον Ιανουάριο του 2014, καθώς και προσωπικά δεδομένα για κάθε μέλος του κοινοβουλίου. Περιέχει δεδομένα σχετικά με τις ημερήσιες συνεδριάσεις, το πρόγραμμα των συζητήσεων, τους λόγους και τις μεταφράσεις τους σε 21 διαφορετικές γλώσσες. Ακόμα περιλαμβάνει πληροφορίες για τους ρόλους των ομιλητών και τα έθνη που εκπροσωπούν, καθώς και τη συμμετοχή των εθνικών κομμάτων, των ευρωπαϊκών κομμάτων και των επιτροπών. Τα δεδομένα είναι ακόμα προσβάσιμα μέσω ενός SPARQL API, στο οποίο δύναται να γίνονται πιο στοχευμένα ερωτήματα, με σκοπό προηγμένες αναλύσεις και λήψη συγκεκριμένων πληροφοριών.

Τα Πρακτικά του Κοινοβουλίου του Καναδά² είναι μια άλλη σημαντική περίπτωση. Ο ιστότοπος παρέχει πρόσβαση στα αρχεία των συζητήσεων που έχουν διεξαχθεί στο Κοινοβούλιο του Καναδά από τις αρχές τις δεκαετίας του 1990 μέχρι και σήμερα. Οι συζητήσεις αυτές καλύπτουν ένα ευρύ φάσμα θεμάτων, επιτρέποντας στους χρήστες να αποκτήσουν εικόνα των νομοθετικών συζητήσεων και των διαδικασιών λήψης αποφάσεων. Φυσικά, μπορούν να αντληθούν πληροφορίες σχετικά με τους ομιλητές που συμμετέχουν στις συζητήσεις, συμπεριλαμβανομένων των ονομάτων τους, των ψήφων τους και των θέσεων τους στο κοινοβούλιο. Οι χρήστες μπορούν να έχουν πρόσβαση στα δεδομένα μέσα από ένα πλήρως φιλικό προς τον χρήστη UI, στο οποίο οι ομιλίες καταγράφονται ανάλογα με το θέμα, την ημερομηνία και τον ομιλητή.

Τέλος από μία άλλη οπτική, υπάρχει μια εργασία η οποία ασχολείται με τα ελληνικά κοινοβουλευτικά δεδομένα και τιτλοφορείται ως "*A Greek Parliament Proceedings Dataset for Computational Linguistics and Political Analysis*"[9]. Είναι μία εργασία που επεξεργάζεται τα αρχεία των συνεδριάσεων του Ελληνικού Κοινοβουλίου, σε μια άλλη διάσταση από αυτή τις παρούσας εργασίας, μιας και δεν εστιάζει στην σημασιολογική διασύνδεση των δεδομένων, αλλά επικεντρώνεται στην αξιολόγηση της ποιότητας του λόγου και στην εξέταση των συναισθημάτων που προκύπτουν από τα λεγόμενα των ομιλητών. Μερικά από αυτά τα αρχεία τα έχουμε εμπλουτίσει για τους σκοπούς της παρούσας εργασίας.

1.4 Οργάνωση κειμένου

Η παρούσα διπλωματική εργασία είναι οργανωμένη σε πέντε κεφάλαια. Στο Κεφάλαιο 1, κάνουμε μια εισαγωγή αναλύοντας το πρόβλημα που σκοπεύουμε να λύσουμε, παρουσιάζοντας και αντίστοιχα παραδείγματα. Στο Κεφάλαιο 2, εμβαθύνουμε στα θεωρητικά τμήματα του σημασιολογικού ιστού και των συναφών τεχνολογιών του εξετάζοντας έννοιες για την κατανόηση των δυνατοτήτων των ανοικτών διασυνδεδεμένων δεδομένων στις κοινοβουλευτικές συζητήσεις. Στο Κεφάλαιο 3, περιγράφουμε την διαδικασία μετατροπής των κειμένων σε δομημένες μορφές, αναλύοντας σε βάθος τα υποσυστήματα που δημιουργήσαμε για τη διαδικασία μετατροπής των κειμένων σε δομημένες μορφές. Στο Κεφάλαιο 4, παρουσιάζουμε μελέτες περιπτώσεων στα τελικά δομημένα αρχεία. Αυτό περιλαμβάνει την εξαγωγή θεματικής μοντελοποίησης και απαντήσεις σε σημασιολογικά ερωτήματα. Τέλος, παρουσιάζουμε στατιστικά δεδομένα που αφορούν τις κοινοβουλευτικές συζητήσεις. Στο

² <https://hansard.opennwt.ca/debates/>

Κεφάλαιο 5, ολοκληρώνουμε την εργασία αναλύοντας τα συμπεράσματα και τις παρατηρήσεις του συστήματος και αναφέρουμε μελλοντικές βελτιώσεις και επεκτάσεις του.

2

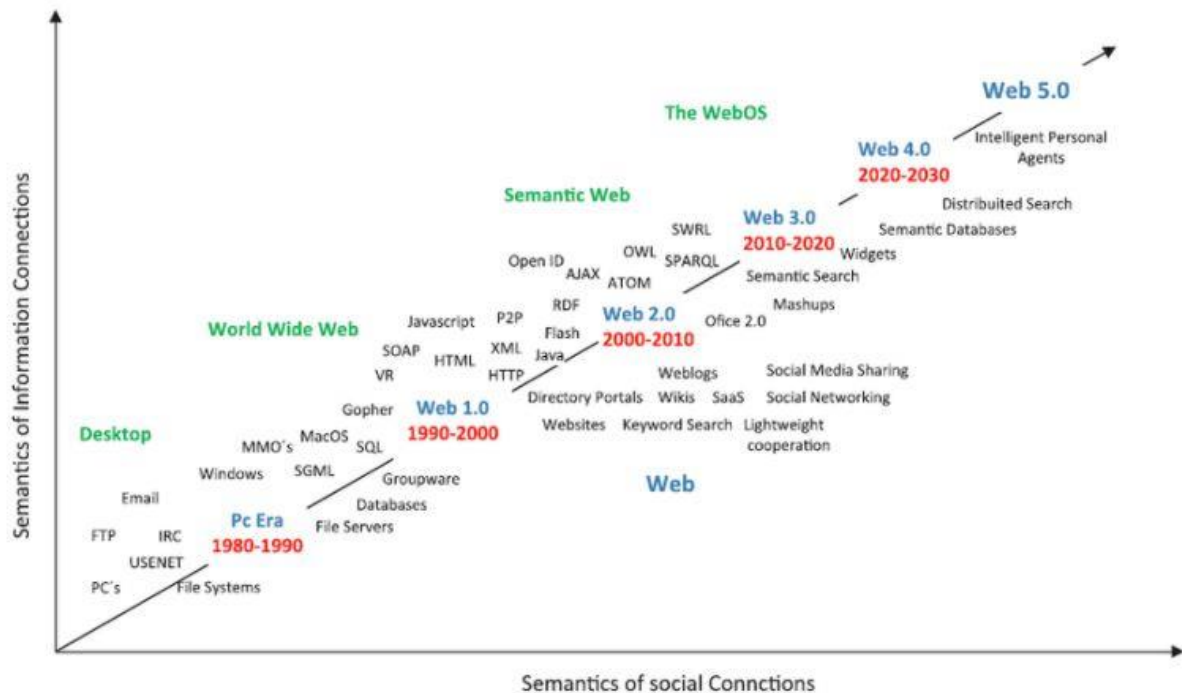
Θεωρητικό υπόβαθρο

Σε αυτό το κεφάλαιο της εργασίας, αναλύουμε τις βασικές θεωρητικές έννοιες που περικλείονται στο νοηματικό περιεχόμενο της εργασίας, και πιο συγκεκριμένα εστιάζουμε στον Σημασιολογικό Ιστό και στα βασικά συστατικά του. Αρχικά, κάνουμε μια παρουσίαση του Σημασιολογικού Ιστού και των Ανοικτών Συνδεδεμένων Δεδομένων. Έπειτα, εξετάζουμε βασικές τεχνολογίες του Σημασιολογικού Ιστού, με έμφαση στο RDF, στην XML, στην πολυμορφία των λεξικών και στα ερωτήματα SPARQL. Στην συνέχεια, κάνουμε μια σύντομη αναφορά στα νομικά ανοικτά συνδεδεμένα δεδομένα, ένα πεδίο μελέτης στο οποίο οι αρχές του Σημασιολογικού Ιστού αλληλεπιδρούν με τις νομικές πληροφορίες. Τέλος, ολοκληρώνουμε παρουσιάζοντας το μοντέλο σημασιολογικής έκφρασης TEI ParlaMint και μιας τεχνικής ανάλυσης κειμένου που χρησιμοποιείται για την αποτελεσματική εξαγωγή τίτλων κειμένου, Latent Dirichlet Allocation (LDA).

2.1 Σημασιολογικός Ιστός – Semantic Web

Η ανάπτυξη του διαδικτύου έχει φέρει επανάσταση στον τρόπο με τον οποίο επικοινωνούμε, έχουμε πρόσβαση σε πληροφορίες και διεξάγουμε επιχειρηματικές δραστηριότητες. Από τα πρώτα βήματά του, ως ένα απλό δίκτυο διασυνδεδεμένων εγγράφων, το διαδίκτυο έχει υποστεί αξιοσημείωτες αλλαγές με την πάροδο των ετών. Μια τέτοια αλλαγή είναι η μετάβαση στον σημασιολογικό ιστό, γεγονός που οδήγησε στη βαθύτερη κατανόηση και νοηματοδότηση του περιεχομένου του ιστού. Αυτή η μεταβολή έχει ανοίξει νέες δυνατότητες για την αναπαράσταση γνώσης, την ολοκλήρωση δεδομένων και την ευφυή αυτοματοποίηση.

Η μετάβαση από τον παραδοσιακό στον σημασιολογικό ιστό αποτελεί σημαντικό ορόσημο στην εξέλιξη του διαδικτύου. Στα πρώτα στάδια του ιστού, οι πληροφορίες παρουσιάζονταν κυρίως σε αδόμητες μορφές, γεγονός που καθιστούσε δύσκολη έως και αδύνατη την αποτελεσματική επεξεργασία και ερμηνεία του περιεχομένου από τις μηχανές. Ωστόσο, καθώς ο ιστός μεγάλωνε σε μέγεθος και πολυπλοκότητα, προέκυψε η ανάγκη για έναν πιο έξυπνο και αποτελεσματικό τρόπο οργάνωσης και κατανόησης του τεράστιου όγκου των διαθέσιμων πληροφοριών. Έτσι γεννιέται η έννοια του σημασιολογικού ιστού, η οποία αποσκοπούσε στο να προσδώσει στο περιεχόμενο του ιστού σαφές νόημα και να επιτρέψει στις μηχανές να κατανοήσουν τα δεδομένα.



Εικόνα 2.1 – Η εξέλιξη του Διαδικτύου³

Η μετάβαση στον σημασιολογικό ιστό φέρνει πολλές υποσχέσεις για διάφορους τομείς. Ενδεικτικά, στον τομέα της ανάκτησης πληροφοριών και των μηχανών αναζήτησης, ο σημασιολογικός ιστός μπορεί να βελτιώσει την ακρίβεια και τη συνάφεια των αποτελεσμάτων αναζήτησης, κατανοώντας το νόημα και το πλαίσιο πίσω από τα ερωτήματα των χρηστών. Σε έναν άλλο τομέα, όπως αυτού του ηλεκτρονικού εμπορίου, ο σημασιολογικός ιστός επιτρέπει πιο έξυπνες συστάσεις προϊόντων και εξατομικευμένες εμπειρίες αγορών με βάση τη βαθύτερη κατανόηση των προτιμήσεων και των αναγκών των πελατών. Τέλος, στην υγειονομική περίθαλψη, ο σημασιολογικός ιστός διευκολύνει τη διαλειτουργικότητα και την διασύνδεση των ιατρικών δεδομένων, οδηγώντας σε βελτιωμένη περίθαλψη των ασθενών, ερευνητική συνεργασία και λήψη αποφάσεων βάσει δεδομένων.

Επιπλέον, ο σημασιολογικός ιστός έχει θετικές επιδράσεις στις εφαρμογές τεχνητής νοημοσύνης (AI) και μηχανικής μάθησης (ML). Παρέχοντας ένα τυποποιημένο πλαίσιο για την αναπαράσταση δεδομένων και την ολοκλήρωση της γνώσης, ο σημασιολογικός ιστός μπορεί να βελτιώσει την εκπαίδευση και την απόδοση των μοντέλων τεχνητής νοημοσύνης, επιτρέποντάς τους να κάνουν πιο τεκμηριωμένες προβλέψεις και συστάσεις. Ανοίγει επίσης το δρόμο για την ανάπτυξη ευφών πρακτόρων και chatbots που μπορούν να κατανοούν και να απαντούν σε ερωτήματα χρηστών με πιο φυσικό και συνειδητό τρόπο.

Κατά συνέπεια, όπως περιγράφεται στην Εικόνα 2.1, η μετάβαση προς τον Σημασιολογικό Ιστό ή τον Ιστό 3.0 είναι ήδη σε εξέλιξη. Η προβλεπόμενη μετάβαση στο Web 4.0, (ή Ευφυής Ιστός), είναι ακόμα σε μεταβατικό στάδιο, όπου αναμένεται να αξιοποιηθεί η τεχνητή νοημοσύνη και η μηχανική μάθηση για να παρέχεται στον χρήστη προσαρμοσμένες υπηρεσίες και περιεχόμενο.

³ Πηγή: <https://myeltcafe.com/articles/evolution-of-web-from-1-0-to-5-0/>

2.2 Ανοιχτά Διασυνδεδεμένα Δεδομένα

Η ιδέα των Ανοικτών Συνδεδεμένων Δεδομένων είναι ένα θεμελιώδες στοιχείο που, εκτός από τον Σημασιολογικό Ιστό, βελτιώνει την αποτελεσματικότητα και τη διαλειτουργικότητα των δεδομένων. Ενθαρρύνοντας την τυποποιημένη και ανοικτή δημοσίευση και σύνδεση δομημένων δεδομένων, τα Ανοιχτά Συνδεδεμένα Δεδομένα προωθούν τα ιδανικά του Σημασιολογικού Ιστού.

Προκειμένου να εκφραστούν τα δεδομένα με τρόπο που να είναι αναγνώσιμο από μηχανήματα, τα Ανοιχτά Συνδεδεμένα Δεδομένα επιτάσσουν τη χρήση ανοικτών προτύπων όπως το RDF (Resource Description Framework). Το RDF προωθεί την τριπλή δομή υποκειμένου-κατηγορουμένου-αντικειμένου των δεδομένων, επιτρέποντας τη δήλωση σύνθετων σημασιολογιών και συνδέσεων εντός των δεδομένων. Αυτό διευκολύνει τον συνδυασμό και την ενσωμάτωση συνόλων δεδομένων από πολλές πηγές, δημιουργώντας ένα δίκτυο γνώσης που είναι πιο εκτεταμένο και συνδεδεμένο.

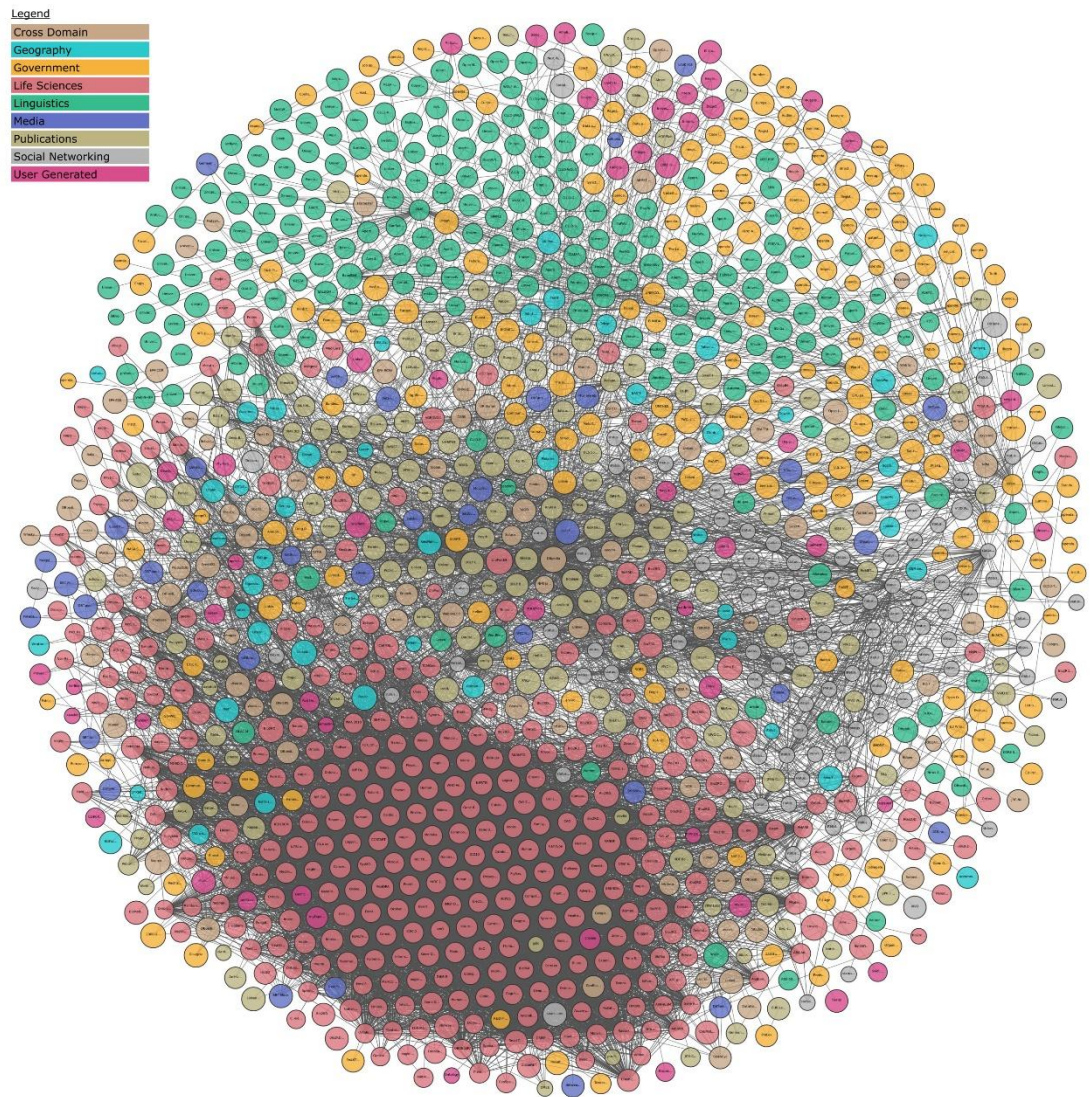
Η χρήση ομοιόμορφων αναγνωριστικών πόρων (URI) για τον ειδικό προσδιορισμό πραγμάτων και εννοιών εντός του οικο-συστήματος των συνδεδεμένων δεδομένων επιβάλλεται από τις αρχές των ανοικτών συνδεδεμένων δεδομένων. Τα URI χρησιμεύουν ως μόνιμα και παγκοσμίως μοναδικά αναγνωριστικά, επιτρέποντας την απρόσκοπτη αναφορά και σύνδεση πόρων σε διαφορετικά σύνολα δεδομένων και τομείς.

Ο Τιμ Μπέρνερς Λι, ο εφευρέτης του Παγκόσμιου Ιστού, περιγράφει τέσσερις βασικές και θεμελιώδεις αρχές για τη δημοσίευση δεδομένων στον Παγκόσμιο Ιστό με τρόπο που όλα τα δημοσιευμένα δεδομένα γίνονται μέρος ενός ενιαίου παγκόσμιου. Πιο συγκεκριμένα:

- Για να διασφαλιστεί ότι κάθε οντότητα ή έννοια έχει μια διακριτή ταυτότητα που την ξεχωρίζει μέσα στο τεράστιο διασυνδεδεμένο δίκτυο δεδομένων, ο πρώτος κανόνας απαιτεί τη χρήση URIs ως ονόματα για τα πράγματα.
- Ο δεύτερος κανόνας καθιστά απλή την αναζήτηση αυτών των ονομάτων με τη χρήση HTTP URIs, επιτρέποντας στους χρήστες να έχουν πρόσβαση και να ανακτούν χρήσιμα δεδομένα που σχετίζονται με τους αναγνωρισμένους πόρους.
- Για την επιτυχή διεύθυνση και αναζήτηση των δεδομένων, ο τρίτος κανόνας δίνει έμφαση στην παρουσίαση σχετικών πληροφοριών κατά την αναζήτηση ενός URI. Για το σκοπό αυτό χρησιμοποιούνται τυποποιημένες τεχνολογίες όπως το RDF και η SPARQL.
- Η συμπερίληψη συνδέσμων προς άλλα URI τονίζεται επίσης από το τέταρτο κριτήριο, διευκολύνοντας την εύρεση νέων σχετικών πόρων και την ανάπτυξη ενός δικτύου διασυνδεδεμένης γνώσης.

Τα συνδεδεμένα δεδομένα λοιπόν είναι μια έννοια που προωθούν τη διασύνδεση και την ενσωμάτωση διαφορετικών συνόλων δεδομένων στο διαδίκτυο, όπως φαίνεται και στην Εικόνα 2.2 (LOD Cloud). Στο διάγραμμα αυτό αποτυπώνεται το μεγαλειώδες δίκτυο των συνδεδεμένων συνόλων δεδομένων, όπου οι κόμβοι αντιπροσωπεύουν μεμονωμένα σύνολα δεδομένων και οι συνδέσεις μεταξύ τους υποδηλώνουν τις σχέσεις τους. Ένα σύνολο δεδομένων, σύμφωνα με τις αρχές που είδαμε παραπάνω, αναπαρίσταται από κάθε κόμβο στο LOD Cloud, το οποίο χρησιμοποιεί μοναδικά αναγνωριστικά (URIs) για τον προσδιορισμό και την αναφορά των πόρων. Οι σύνδεσμοι και οι συνδέσεις μεταξύ των διαφόρων συνόλων δεδομένων αναπαρίστανται από τους συνδέσμους μεταξύ των κόμβων, επιτρέποντας την

εύκολη πλοήγηση και την εξερεύνηση πληροφοριών. Το διάγραμμα της εικόνας δείχνει τη δύναμη της διασύνδεσης δεδομένων μεταξύ τους, επιτρέποντας στους χρήστες να περιηγηθούν μεταξύ συνόλων δεδομένων, να βρουν νέα δεδομένα και να αποκτήσουν ολοκληρωμένες γνώσεις, αξιοποιώντας τη συλλογική σοφία που υπάρχει στο οικοσύστημα των συνδεδεμένων δεδομένων.

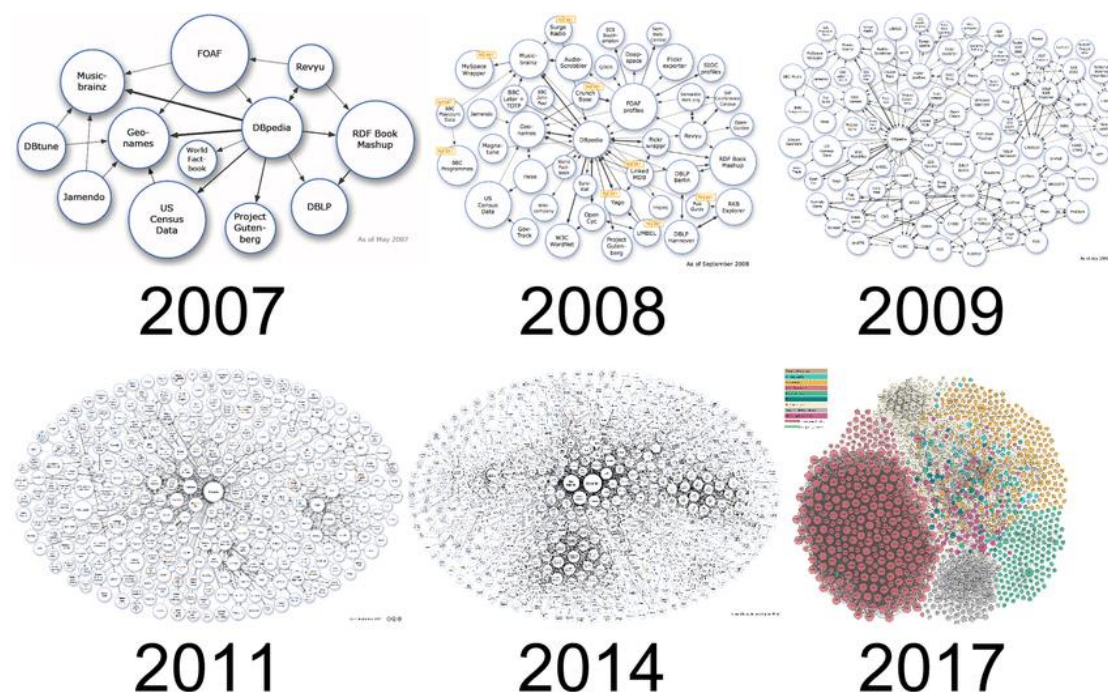


Εικόνα 2.2 – The Linked Open Data Cloud⁴

Στο γράφημα της Εικόνας 2.3 απεικονίζεται η εκθετική ανάπτυξη των Συνδεδεμένων Ανοικτών Δεδομένων (LOD) από το 2007 έως και τα τελευταία χρόνια, αναδεικνύοντας την αξιοσημείωτη πρόοδο που έχουν σημειώσει τα Διασυνδεδεμένα Δεδομένα με την πάροδο του χρόνου. Το γράφημα καταγράφει το αυξανόμενο δίκτυο πόρων συνδεδεμένων δεδομένων, δείχνοντας την αύξηση του αριθμού των συνόλων δεδομένων που δημοσιεύονται ως LOD. Η επέκταση αυτή αντανακλά την αυξανόμενη αποδοχή και εκτίμηση των αρχών των συνδεδεμένων δεδομένων σε ένα ευρύ φάσμα πεδίων και τομέων. Η ποσότητα και η ποικιλία

⁴ Πηγή: <http://cas.lod-cloud.net/>

των διασυνδεδεμένων συνόλων δεδομένων έχουν αυξηθεί τρομερά καθώς όλο και περισσότερες επιχειρήσεις, κοινωνίες και κυβερνήσεις υιοθετούν την ιδέα των συνδεδεμένων δεδομένων. Η ανάπτυξη ενός τεράστιου, διασυνδεδεμένου γράφου γνώσης ως αποτέλεσμα αυτής της ανάπτυξης κατέστησε δυνατή την καλύτερη ενσωμάτωση, ανακάλυψη και χρήση δεδομένων.



Εικόνα 2.3 – Χρονική ανάπτυξη των συνδεδεμένων ανοικτών δεδομένων⁵

2.3 Τεχνολογίες Υποστήριξης Δεδομένων

Στον τομέα του Σημασιολογικού Ιστού και των Συνδεδεμένων Δεδομένων περιλαμβάνονται διάφορες τεχνολογίες για την αποτελεσματική αναπαράσταση, διασύνδεση, και διαλειτουργικότητα των δεδομένων. Οι τεχνολογίες αυτές προσφέρουν το πλαίσιο για την οργάνωση, τη σύνδεση και τη συλλογή γνώσης από διάφορα σύνολα δεδομένων. Μια ευέλικτη και ευρέως χρησιμοποιούμενη γλώσσα για την κωδικοποίηση δομημένων δεδομένων, η XML (eXtensible Markup Language) προωθεί την ανταλλαγή δεδομένων και τη διαλειτουργικότητα. Προκειμένου να καταστεί δυνατή η ανάπτυξη πλούσιων σημασιολογικών αναπαραστάσεων, το RDF (Resource Description Framework) παρουσιάζει ένα τυποποιημένο μοντέλο για την περιγραφή και τη σύνδεση δεδομένων. Ο πυρήνας των συνδεδεμένων δεδομένων, το RDF προσφέρει έναν αποτελεσματικό τρόπο περιγραφής σχέσεων, ιδιοτήτων και πληροφοριών. Με τη βοήθεια της ισχυρής γλώσσας ερωτημάτων SPARQL, οι χρήστες μπορούν να λάβουν συγκεκριμένα δεδομένα, να πραγματοποιήσουν περίπλοκες συνδέσεις και να αποκτήσουν κατανόηση από τα διασυνδεδεμένα δεδομένα. Αυτές οι τεχνολογίες συνεργάζονται για να δημιουργήσουν μια πλήρη εργαλειοθήκη που επιτρέπει στους ερευνητές, τους

⁵ Πηγή: https://www.researchgate.net/figure/Growth-of-Linked-Open-Data-since-2007-1-The-amount-of-data-sets-published-as-LOD-have_fig2_331748480

προγραμματιστές και τους επαγγελματίες των δεδομένων να αξιοποιούν πλήρως τα έργα του σημασιολογικού ιστού και των συνδεδεμένων δεδομένων.

2.3.1 Extensible Markup Language (XML)

Η eXtensible Markup Language, γνωστή ως XML, είναι μια ευρέως υιοθετημένη τεχνολογία στον τομέα του Σημασιολογικού Ιστού και των συνδεδεμένων δεδομένων, καθώς είναι μια ισχυρή γλώσσα σήμανσης που έχει επηρεάσει σημαντικά την ανάπτυξη της ανταλλαγής πληροφοριών στο διαδίκτυο. Η αφηρητή της εντοπίζεται στις αρχές της δεκαετίας του 1970, όταν κατέστη για πρώτη φορά αναγκαίος ένας δομημένος τρόπος αναπαράστασης και ανταλλαγής δεδομένων.

Στις δεκαετίες του 1980 και 1990 δημιουργήθηκαν διάφορες γλώσσες σήμανσης (όπως η HTML), η καθεμία με μοναδικούς περιορισμούς και προβλεπόμενες χρήσεις. Ωστόσο, για να αντιμετωπιστούν οι αυξανόμενες απαιτήσεις για ανταλλαγή δεδομένων σε διάφορες πλατφόρμες και συστήματα, απαιτήθηκε μια πιο προσαρμόσιμη και επεκτάσιμη λύση, με αποτέλεσμα στα τέλη της δεκαετίας του 1990, να αναπτυχθεί η XML.

Η Κοινοπραξία του Παγκόσμιου Ιστού⁶ (W3C) ξεκίνησε τη διαδικασία τυποποίησης της XML το 1996 και καθόρισε τη σύνταξη και τη σημασιολογία της. Ο κύριος στόχος ήταν η ανάπτυξη μιας γλώσσας που θα επέτρεπε στους χρήστες να κατασκευάζουν τις δικές τους μοναδικές γλώσσες σήμανσης, καθιστώντας την αρκετά ευέλικτη ώστε να μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα εφαρμογών και δομών δεδομένων.

Η κύρια καινοτομία της XML είναι το πόσο απλή και κατανοητή είναι. Τα στοιχεία περιέχονται σε αγκύλες (<) / (>) και διατάσσονται ιεραρχικά χρησιμοποιώντας μια μέθοδο βασισμένη σε ετικέτες. Τόσο οι άνθρωποι όσο και οι μηχανές μπορούν να ερμηνεύσουν τα δεδομένα, δεδομένου ότι αυτές οι ετικέτες καθορίζουν τη δομή και το νόημά τους.

Παρακάτω παρατίθεται ένα απόσπασμα από μια πιθανή απεικόνιση της XML με χρήση ετικετών για να γίνει κατανοητή η σύνταξη της γλώσσας:

```
<person>
  <name>John Papanikolaou</name>
  <age>23</age>
  <address>
    <street>Main Street</street>
    <city>Athens</city>
    <country>Greece</country>
  </address>
</person>
```

Στο παράδειγμα αυτό, οι ετικέτες XML ενσωματώνουν διαφορετικά κομμάτια πληροφοριών για ένα άτομο. Το βασικό στοιχείο (root) είναι η ετικέτα "person", η οποία περιέχει φωλιασμένες ετικέτες όπως "name", "age" και "address". Κάθε ετικέτα αντιπροσωπεύει ένα μοναδικό κομμάτι δεδομένων. Για παράδειγμα, οι ετικέτες "name" και "age" περιέχουν το

⁶ <https://www.w3.org/>

όνομα και την ηλικία του ατόμου αντίστοιχα, ενώ η ετικέτα "address" έχει φωλιασμένες τις ετικέτες "street", "city" και "country" για να αντικατοπτρίζει τις πληροφορίες διεύθυνσης.

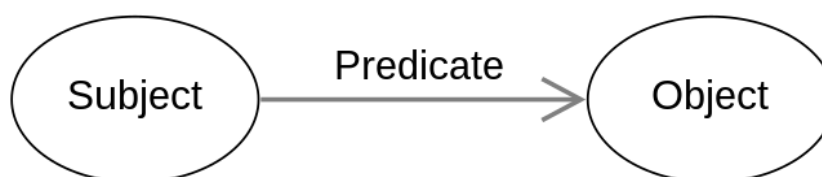
Στο πλαίσιο των συνδεδεμένων δεδομένων, η XML αποτελεί βασική τεχνική για τη δόμηση και τη δημιουργία εγγράφων αναγνώσιμων από μηχανήματα για μη δομημένα κείμενα. Αυτά τα έγγραφα μπορούν να επεξεργαστούν και να επισημανθούν με χρήσιμα στοιχεία και χαρακτηριστικά με τη χρήση XML, με αποτέλεσμα την παραγωγή καλά διαμορφωμένων εγγράφων XML. Τα επόμενα βήματα της επεξεργασίας, της διασύνδεσης και της αναζήτησης δεδομένων με τη χρήση διαφόρων τεχνολογιών σημασιολογικού ιστού καθίστανται δυνατά με αυτή τη διαδικασία.

Σε καταστάσεις που αφορούν την ενσωμάτωση δεδομένων, όπου διάφορες πηγές πρέπει να εναρμονιστούν σε μια ενιαία αναπαράσταση, η XML είναι επίσης απαραίτητη. Είναι απλούστερη η αντιστοίχιση και ο μετασχηματισμός διαφορετικών συνόλων δεδομένων σε μια ενιαία μορφή λόγω της καθολικής σύνταξης της XML για την αναπαράσταση δομών δεδομένων.

2.3.2 Resource Description Framework (RDF)

Το Resource Description Framework (RDF) χρησιμεύει ως μοντέλο για την αναπαράσταση και τη σύνδεση δεδομένων. Το RDF παρέχει μια δομημένη και ευέλικτη προσέγγιση για την περιγραφή των πόρων, των χαρακτηριστικών τους και των μεταξύ τους σχέσεων.

Οι τριπλέτες υποκειμένου-προγνωστικού-αντικειμένου (*subject – predicate – object*) είναι τα θεμελιώδη δομικά στοιχεία των αναπαραστάσεων πληροφοριών RDF (εικόνα 2.4). Οι πόροι που περιγράφονται αντιπροσωπεύονται από το υποκείμενο, το κατηγορήμα και το αντικείμενο, το οποίο αντιπροσωπεύει την τιμή ή έναν άλλο πόρο στον οποίο παραπέμπει η ιδιότητα. Αυτές οι τριπλέτες μπορούν να χρησιμοποιηθούν για τη δημιουργία ενός γραφήματος συνδεδεμένων δεδομένων, το οποίο λειτουργεί ως βάση για έναν Knowledge Graph. Knowledge Graph, ή αλλιώς «γράφος γνώσης» είναι μια βάση γνώσης που χρησιμοποιεί ένα μοντέλο ή μια τοπολογία δεδομένων με δομή γράφου για την ενσωμάτωση δεδομένων.



Εικόνα 2.4 – Βασική σύνταξη RDF⁷

Το RDF ωφελεί τον Σημασιολογικό Ιστό με διάφορους τρόπους. Πρώτον, προσφέροντας ένα συνεπές μοντέλο δεδομένων και λεξιλόγιο για την κωδικοποίηση πληροφοριών, διευκολύνει τη συγχώνευση δεδομένων από πολλές πηγές. Διαφορετικά σύνολα δεδομένων μπορούν να ενσωματωθούν, να συνδεθούν και να αναζητηθούν συλλογικά χάρη στη διαλειτουργικότητα.

Χρησιμοποιώντας οντολογίες, το RDF διευκολύνει την επεκτασιμότητα. Οι έννοιες, οι συνδέσεις, οι περιορισμοί και γενικότερα η γνώση που αφορά έναν συγκεκριμένο τομέα

⁷Πηγή: https://en.wikipedia.org/wiki/Resource_Description_Framework#Statement_reification_and_context

αναπαρίστανται τυπικά και δομικά με οντολογίες. Φυσικά, αυτές οι οντολογίες μπορούν να επεκταθούν όσο αναπτύσσονται νέες ιδέες και πληροφορίες, δίνοντας στα RDF την ευελιξία να προσαρμόζεται στις μεταβαλλόμενες απαιτήσεις του κλάδου.

Επιπλέον, το RDF διευκολύνει την έκφραση σχέσεων και σημασιολογίας μέσα στα δεδομένα. Το RDF παρέχει μοναδικά και μόνιμα αναγνωριστικά για τους πόρους, χρησιμοποιώντας για αυτό το σκοπό τα URI. Αυτό επιτρέπει τη σωστή αναφορά και σύνδεση των δεδομένων σε όλα τα σύνολα δεδομένων. Η ικανότητα σύνδεσης των σχετικών πηγών μεταξύ τους βελτιώνει την κατανόηση και την ερμηνεία των δεδομένων.

Τέλος, τα δεδομένα RDF μπορούν να αποθηκευτούν και να αναζητηθούν αποτελεσματικά με τη χρήση ενός συστήματος διαχείρισης τριπλετών RDF. Ένα τέτοιο είναι το Apache Jena Fuseki, μια βάση δεδομένων RDF ανοικτού κώδικα. Η ικανότητα του Apache Fuseki να αποθηκεύει δεδομένα RDF με ιδιαίτερα κλιμακούμενο και αποτελεσματικό τρόπο είναι ένα από τα χαρακτηριστικά που το διακρίνουν. Χρησιμοποιεί το πλαίσιο Apache Jena, το οποίο παρέχει πλήρεις δυνατότητες διαχείρισης δεδομένων RDF, συμπεριλαμβανομένης της αποθήκευσης δεδομένων και της βελτιστοποίησης ερωτημάτων.

Επιπλέον, το Apache Fuseki προσφέρει ένα HTTP SPARQL endpoint που επιτρέπει στους χρήστες να αλληλεπιδρούν με τα αποθηκευμένα δεδομένα RDF, χρησιμοποιώντας τη γλώσσα ερωτημάτων SPARQL. Με τη βοήθεια της SPARQL, οι χρήστες μπορούν να εκτελούν εξελιγμένες λειτουργίες σύνδεσης, να λαμβάνουν συγκεκριμένα δεδομένα και να εξετάζουν τις αλληλένδετες σχέσεις που περιλαμβάνονται στα σύνολα δεδομένων. Η εκτέλεση ερωτημάτων SPARQL στην υποκείμενη βάση δεδομένων RDF γίνεται απλή και τυποποιημένη από το τελικό σημείο SPARQL που προσφέρει το Fuseki.

2.3.3 Λεξιλόγια Σημασιολογικού Ιστού

Τα λεξιλόγια του Σημασιολογικού Ιστού, γνωστά και ως οντολογίες, διαδραματίζουν κρίσιμο ρόλο στην αναπαράσταση και οργάνωση των δεδομένων στο οικοσύστημα του Σημασιολογικού Ιστού. Αυτά τα λεξιλόγια παρέχουν ένα κοινό και τυποποιημένο σύνολο όρων, σχέσεων και περιορισμών που επιτρέπουν τη δομημένη περιγραφή και ερμηνεία των δεδομένων.

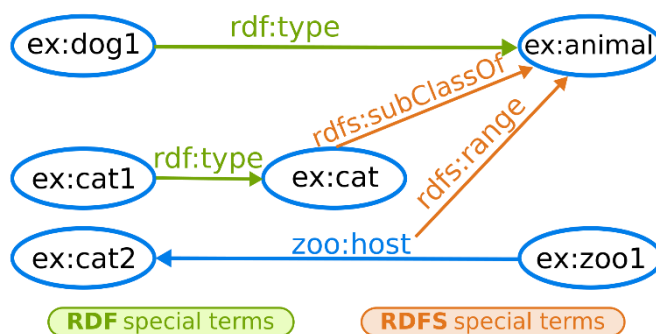
Προορίζονται για την καταγραφή των ιδεών και των συνδέσεων που είναι μοναδικές σε έναν συγκεκριμένο τομέα. Παρέχουν στη γνώση μια τυπική αναπαράσταση, επιτρέποντας τη μοντελοποίηση και την έκφραση της σημασιολογίας που αφορά ένα συγκεκριμένο θεματικό κύκλο. Τα λεξιλόγια ακόμα επιτρέπουν την ανταλλαγή και την ενσωμάτωση δεδομένων σε διάφορες εφαρμογές και τομείς περιγράφοντας ιδέες και τις σχέσεις τους.

Η ικανότητα του Σημασιολογικού Ιστού να περιγράφει πολλές μορφές πληροφοριών με τυποποιημένο και διαλειτουργικό τρόπο διευκολύνεται από διάφορα λεξιλόγια, καθένα από τα οποία έχει ξεχωριστό σκοπό και σύνολο ορολογίας. Οι εφαρμογές και τα συστήματα που χρησιμοποιούν αυτά τα λεξιλόγια μπορούν να επωφεληθούν από το σύνολο της γνώσης που αντιπροσωπεύουν συλλογικά, επιτρέποντας βαθύτερη ενσωμάτωση δεδομένων, αποτελεσματικότερη αναζήτηση και βελτιωμένη σημασιολογική διαλειτουργικότητα.

Τα λεξιλόγια του Σημασιολογικού Ιστού έχουν πολλά πλεονεκτήματα. Για αρχή, επιτρέπουν την διασύνδεση δεδομένων παρέχοντας μια κοινή κατανόηση των δεδομένων σε πολλές εφαρμογές και συστήματα. Τα λεξιλόγια καθιστούν δυνατή την ουσιαστική και κατανοητή αναπαράσταση των δεδομένων από τις μηχανές, επιτρέποντας τη βελτιωμένη

αναζήτηση, ανάκτηση και ανεύρεση πληροφοριών. Παρέχουν τέλος εξελιγμένη εξαγωγή συμπερασμάτων διευκολύνοντας την αυτόματη παραγωγή νέων πληροφοριών.

Πιο συγκεκριμένα, ένα από τα ευρέως χρησιμοποιούμενα λεξιλόγια του Σημασιολογικού Ιστού είναι το RDF Schema (RDFS), το οποίο είναι ένα σύνολο κανόνων που επεκτείνουν τις δυνατότητες του RDF. Το RDFS παρέχει τα θεμέλια για τον ορισμό κλάσεων, ιδιοτήτων και σχέσεων μεταξύ τους, επιτρέποντας την ιεραρχική ταξινόμηση και τον προσδιορισμό ιδιοτήτων άλλων πόρων, όπως οι τομείς και τα εύρη των ιδιοτήτων. Η ικανότητά του να δημιουργεί κλάσεις και ιδιότητες με ιεραρχικό τρόπο, η οποία επιτρέπει την ανάπτυξη οντολογιών στις οποίες οι κλάσεις μπορούν να κληρονομούν ιδιότητες από τις γονικές τους κλάσεις, είναι ένα από τα σημαντικότερα επιτεύγματά του.



Εικόνα 2.5 – RDF(S) παράδειγμα ⁸

Υπάρχουν διάφορα ειδικά λεξιλόγια για διάφορους τομείς και κλάδους, εκτός του RDFS. Ακολουθούν ορισμένα από τα πιο συνηθισμένα λεξιλόγια του Σημασιολογικού Ιστού:

- FOAF (Friend of a Friend)⁹

Το FOAF είναι ένα λεξιλόγιο που χρησιμοποιείται για τον χαρακτηρισμό των ανθρώπων και των συνδέσεών τους. Δίνει ορισμούς για όρους που δηλώνουν προσωπικά δεδομένα όπως ονόματα, διευθύνσεις ηλεκτρονικού ταχυδρομείου και προφίλ στα μέσα κοινωνικής δικτύωσης. Το FOAF επιτρέπει την απεικόνιση των διαπροσωπικών δεσμών, όπως των φιλικών σχέσεων, των συνεργασιών και της ομαδικής εργασίας.

- SKOS (Simple Knowledge Organization System) ¹⁰

Το SKOS είναι μία οντολογία που επικεντρώνεται στην περιγραφή συστημάτων οργάνωσης γνώσης, όπως ταξινομίες (taxonomies), θησαυροί (thesauri), και συστήματα ταξινόμησης.

- Dublin Core¹¹

Για την περιγραφή των στοιχείων μεταδεδομένων των ψηφιακών πόρων, το Dublin Core είναι ένα ευρέως χρησιμοποιούμενο λεξιλόγιο. Παρέχει όρους για τη συλλογή θεμελιωδών λεπτομερειών σχετικά με τους πόρους, συμπεριλαμβανομένων των ονομάτων, των συγγραφέων, των θεμάτων και των ημερομηνιών.

⁸ Πηγή: https://en.wikipedia.org/wiki/RDF_Schema#RDFS_entailment

⁹ <http://xmlns.com/foaf/0.1/>

¹⁰ <https://www.w3.org/TR/skos-reference/>

¹¹ <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

- Schema.org¹²

Σε μια προσπάθεια να παρέχουν ένα κοινό λεξιλόγιο για δομημένα δεδομένα στον ιστό, οι μεγάλες μηχανές αναζήτησης, συμπεριλαμβανομένων των Google, Microsoft και Yahoo, συνεργάστηκαν για τη δημιουργία του Schema.org. Παρέχει μια μεγάλη ποικιλία ορολογίας για την κατηγοριοποίηση πραγμάτων όπως εταιρείες, αντικείμενα, γεγονότα και σχέσεις σε πολλούς διαφορετικούς τομείς. Η βελτίωση στην ποιότητα και την ποσότητα στα αποτελέσματα των μηχανών αναζήτησης καθίστανται δυνατή λόγω αυτού του λεξιλογίου.

- DBpedia¹³

Μια δημοφιλής και εκτεταμένη οντολογία που αναπαριστά τη γνώση που λαμβάνεται από τη Βικιπαίδεια με οργανωμένο τρόπο. Προσφέρει ένα ευρύ φάσμα κλάσεων, χαρακτηριστικών και συνδέσεων που αντιπροσωπεύουν τα διάφορα πεδία και θέματα που καλύπτονται στα άρθρα της Wikipedia.

Τέλος αξίζει να αναφέρουμε πως η Web Ontology Language (OWL) θεωρείται το καθιερωμένο πρότυπο για την αναπαράσταση οντολογιών στον Σημασιολογικό Ιστό. Επιτρέπει τη λεπτομερή αναπαράσταση της γνώσης που αφορά διάφορα αντικείμενα, ομάδες αντικειμένων και τις μεταξύ τους συνδέσεις. Η OWL είναι μια γλώσσα βασισμένη στην υπολογιστική λογική, έτσι ώστε η γνώση που εκφράζεται στην OWL να μπορεί να αξιοποιηθεί από προγράμματα υπολογιστών. Μερικά γνωστά κατηγορήματα αυτής της γλώσσας είναι owl:Ontology, owl:sameAs κτλ.

2.3.4 SPARQL

Η SPARQL¹⁴ είναι μια γλώσσα ερωτημάτων ειδικά σχεδιασμένη για την αναζήτηση δεδομένων RDF (Resource Description Framework) στον Σημασιολογικό Ιστό. Παρέχει ένα τυποποιημένο και εκφραστικό συντακτικό για την ανάκτηση και τον χειρισμό πληροφοριών από RDF δεδομένα.

Τα ερωτήματα SPARQL ακολουθούν μια παραπλήσια δομή σύνταξης όπως και τα απλά ερωτήματα SQL, δηλαδή ακολουθούν το μοτίβο *SELECT-WHERE-FILTER*. Πιο συγκεκριμένα, στη συνθήκη *SELECT* ορίζουμε μεταβλητές που επιστρέφονται στα αποτελέσματα του ερωτήματος. Στο *WHERE* καθορίζονται τα μοτίβα και οι απαιτήσεις για την αντιστοίχιση τριπλών RDF στο σύνολο δεδομένων, ενώ στην προαιρετική συνθήκη *FILTER* δίνονται οι απαιτήσεις για το φιλτράρισμα των αποτελεσμάτων βάσει συγκεκριμένων κριτηρίων. Βέβαια, η μόνη βασική διαφορά που αξίζει να αναφερθεί μεταξύ της σύνταξης της SPARQL και της SQL έγκειται στην αντιμετώπιση των προθεμάτων (prefix). Τα προθέματα χρησιμοποιούνται στην SPARQL για τον ορισμό συντομογραφιών των οντολογιών, καθιστώντας το ερώτημα συντομότερο και πιο κατανοητό. Οι χρήστες μπορούν να δώσουν σε ένα URI μιας οντολογίας-λεξικού ένα γρήγορο πρόθεμα χρησιμοποιώντας τη λέξη-κλειδί

¹² <https://schema.org/>

¹³ <https://www.dbpedia.org/>

¹⁴ <https://www.w3.org/TR/sparql11-overview/>

PREFIX και στη συνέχεια να χρησιμοποιήσουν αυτό το πρόθεμα για να αναφερθούν σε πόρους ή χαρακτηριστικά εντός αυτού του χώρου ονομάτων σε όλη τη διάρκεια του ερωτήματος.

Ένα απλό παράδειγμα φαίνεται στην εικόνα 2.6, όπου είναι σχεδιασμένο για την εξαγωγή ονομάτων FOAF και των αντίστοιχων πόρων τους από ένα σύνολο δεδομένων. Χρησιμοποιώντας ως prefix οντολογία την FOAF, αναζητούνται πόροι (?x) που έχουν την ιδιότητα FOAF name (?name). Τα αποτελέσματα αποτελούνται από ζεύγη το ?x και το ?name, όπου αντιπροσωπεύει τον πόρο και το όνομα FOAF που σχετίζεται με αυτόν τον πόρο, αντίστοιχα.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
```

```
SELECT ?x ?name
```

```
WHERE { ?x foaf:name ?name }
```

Εικόνα 2.6 – Ενδεικτικό απλό ερώτημα σε γλώσσα Sparql

Η SPARQL διαδραματίζει πρωταγωνιστικό ρόλο στο Apache Jena Fuseki, όπως είδαμε και νωρίτερα. Στο σύνολο δεδομένων RDF που είναι αποθηκευμένα στο Fuseki, οι χρήστες μπορούν να δημιουργήσουν ερωτήματα SPARQL για να αποκτήσουν τα απαραίτητα δεδομένα και να εκτελέσουν περίπλοκες λειτουργίες. Οι χρήστες μπορούν να εξερευνήσουν τον γράφο του RDF, να φιλτράρουν τα αποτελέσματα και να αθροίσουν τα δεδομένα σύμφωνα με τις μοναδικές τους ανάγκες χάρη στην εκφραστική φύση της SPARQL.

2.4 Ανοικτά Κυβερνητικά Δεδομένα – LegalDocML

Τα διασυνδεδεμένα ανοικτά κυβερνητικά δεδομένα (LOGD), είναι μια «ιδέα» που προωθεί την δημοσιοποίηση, τη διαλειτουργικότητα και την ενσωμάτωση των κυβερνητικών δεδομένων στο διαδίκτυο. Εφαρμόζει τις αρχές των Συνδεδεμένων Ανοικτών Δεδομένων στο πεδίο της κυβέρνησης, αποσκοπώντας να μεγιστοποιήσει το πλέγμα των ανοικτών δεδομένων και να προωθήσει την καινοτομία, τη συνεργασία και τη διαφάνεια.

Τα LOGD περιλαμβάνουν την κοινή χρήση δημόσιων δεδομένων σε ανοικτά και τυποποιημένα αρχεία, τηρώντας παράλληλα τις αρχές των Συνδεδεμένων Δεδομένων. Προκειμένου να προσδιοριστούν οι πόροι και οι σχέσεις, τα δεδομένα διατίθενται ως RDF τριπλέτες, τα οποία χρησιμοποιούν τα μοναδικά αναγνωριστικά (URIs). Τα κυβερνητικά δεδομένα ορίζονται ως διασυνδεδεμένα όταν ακολουθούνται τις αρχές των ανοικτών δεδομένων, καθιστώντας παράλληλα απλή τη σύνδεση και την ανάμειξή τους με άλλα σύνολα δεδομένων για τη δημιουργία ενός πλούσιου ιστού διασυνδεδεμένης γνώσης.

Η υιοθέτηση των LOGD έχει μια σειρά από πλεονεκτήματα. Πρώτα απ' όλα, διευκολύνει την πρόσβαση και τη χρήση κυβερνητικών δεδομένων από πολίτες, ακαδημαϊκούς, εταιρείες και πολιτικούς. Παρέχοντας ανοικτά και οργανωμένα δεδομένα, το LOGD επιτρέπει σε άτομα και οργανισμούς να εξετάζουν, να αξιολογούν και να εξάγουν συμπεράσματα από τα δημόσια δεδομένα.

Δεύτερον, τα LOGD ενθαρρύνουν τη λογοδοσία και την διαφάνεια στις κυβερνητικές δραστηριότητες. Οι κυβερνήσεις μπορούν να καλλιεργήσουν την εμπιστοσύνη και να δώσουν τη δυνατότητα στους πολίτες να εξετάζουν τα γεγονότα, καθιστώντας τα δεδομένα άμεσα διαθέσιμα. Τα LOGD διευκολύνουν την παρακολούθηση των δημόσιων υπηρεσιών, την

ενθάρρυνση της συμμετοχής των πολιτών και την καλύτερη κατανόηση των κυβερνητικών διαδικασιών και νόμων.

Επιπλέον, τα ανοικτά κυβερνητικά δεδομένα προωθούν τη συνεργασία και την ανταλλαγή πληροφοριών μεταξύ των κυβερνήσεων και σε πολλούς τομείς. Τα κυβερνητικά δεδομένα μπορούν να συνδεθούν με άλλα σύνολα δεδομένων για την δημιουργία νέων σχέσεων και γνώσεων που θα επιτρέψουν τη διατομεακή ανάλυση και πιθανή επίλυση προβλημάτων.

Κυβερνήσεις σε όλο τον κόσμο έχουν προχωρήσει στην «ανοικτή» δημοσίευση των τοπικών τους κυβερνητικών αρχείων, παρέχοντας ποικίλες πληροφορίες, συμπεριλαμβανομένων στατιστικών στοιχείων για τον πληθυσμό, τις δαπάνες, τους νόμους και τα ψηφίσματα. Οι χρήστες μπορούν να έχουν πρόσβαση στα δεδομένα και να τα αναζητούν με τυποποιημένο και διαλειτουργικό τρόπο χάρη στις εξειδικευμένες πύλες, τα API και τα SPARQL endpoints μέσω των οποίων διατίθενται αυτά τα σύνολα δεδομένων.

Η αναπαράσταση και η διαχείριση νομοθετικών και νομικών εγγράφων επιτυγχάνεται με ποικίλα πρότυπα. Ένα βασικό είναι το LegalDocML¹⁵, παλιά γνωστό ως Akoma Ntoso, το οποίο είναι ένα πρότυπο βασισμένο στην XML που έχει αναπτυχθεί και συντηρείται από τον οργανισμό *Oasis Open*¹⁶. Το LegalDocML δημιουργεί έναν δομημένο μορφότυπο για την απεικόνιση κοινοβουλευτικών, νομοθετικών και δικαστικών εγγράφων που υποστηρίζει τη διαλειτουργικότητα, την προσβασιμότητα και τη σημασιολογία, βασιζόμενο στις ιδέες των Συνδεδεμένων Ανοικτών Δεδομένων.

Πιο συγκεκριμένα, το LegalDocML παρουσιάζει μια ενδεδειγμένη συλλογή στοιχείων και χαρακτηριστικών που είναι ειδικά σχεδιασμένα για την αποτύπωση της μορφής και του περιεχομένου των δημόσιων επίσημων κειμένων. Περιλαμβάνει ένα εύρος χαρακτηριστικών που παρατηρούνται σε νομικά κείμενα, όπως επικεφαλίδες, υποκεφαλίδες, άρθρα, παραγράφους και παραπομπές. Τα νομικά έγγραφα μπορούν να αναπαρασταθούν με ομοιόμορφο και συνεπή τρόπο συμμορφούμενα με το πρότυπο LegalDocML, απλοποιώντας την ανταλλαγή, την ανάλυση και την επαναχρησιμοποίησή τους.

Πρόσθετα οφέλη προκύπτουν από την ενσωμάτωση του LegalDocML με τα Συνδεδεμένα Ανοικτά Κυβερνητικά Δεδομένα. Ως μια αυστηρή και τυποποιημένη μορφή παρουσίασης των δεδομένων, η διασύνδεση των εγγράφων καθιστά δυνατή με άλλα συναφή δεδομένα και σύνολα δεδομένων, αξιοποιώντας τις έννοιες των Συνδεδεμένων Δεδομένων. Με τη χρήση αυτής της δυνατότητας, τέτοιου είδους κείμενα μπορούν να ενσωματωθούν απρόσκοπτα με μεταδεδομένα, νομοθετικές και δικαστικές αποφάσεις, και άλλες σχετικές πηγές δεδομένων, δημιουργώντας ένα πιο λεπτομερές και ολοκληρωμένο γράφημα πολιτικής γνώσης.

Το LegalDocML χρησιμοποιείται τόσο για την βελτίωση της προσβασιμότητας των κοινοβουλευτικών κειμένων, όσο και για την μηχανική τους αναγνώριση – ανάγνωση. Η προηγμένη αναζήτηση, η σημασιολογική ανάλυση και η αυτοματοποιημένη επεξεργασία των νομοθετικών δεδομένων καθίστανται δυνατές χάρη στη δομημένη αναπαράσταση των κειμένων στο LegalDocML.

Κυβερνήσεις, δικαστήρια και άλλοι οργανισμοί σε όλο τον κόσμο έχουν υιοθετήσει ευρέως το LegalDocML ως πρότυπο για την παρουσίαση νομοθετικών κειμένων. Η χρήση του

¹⁵ <https://www.oasis-open.org/committees/legaldocml/>

¹⁶ <https://www.oasis-open.org/>

διασφαλίζει ότι οι κυβερνητικές πληροφορίες είναι τυποποιημένες και διαλειτουργικές, προωθώντας τη διεθνή συνεργασία, τη συγκριτική νομοθετική έρευνα και τη δημιουργία κυβερνητικών οντολογιών και βάσεων γνώσης.

2.5 *TEI ParlaMint*

Η κοινοπραξία Text Encoding Initiative (TEI)¹⁷, στην ανάγκη να τυποποιήσει την παρουσίαση του κειμενικού υλικού, δημιούργησε και συντηρεί ένα σύνολο κανόνων για την κωδικοποίηση ψηφιακού κειμένου. Το TEI προσφέρει ένα πλαίσιο για την μηχανική παραγωγή αναγνώσιμων αναπαραστάσεων κειμένων, ειδικά για ακαδημαϊκούς και μελετητές που εργάζονται με ψηφιακά κείμενα, ιδίως στις ανθρωπιστικές επιστήμες,

Η ομάδα CLARIN¹⁸, βασιζόμενη στους κανόνες που έθεσε η TEI, ανέπτυξε ένα νέο πρότυπο για την έκφραση δεδομένων εξειδικευμένο στον τομέα του κοινοβουλευτικού κλάδου, με χαρακτηριστική ονομασία «ParlaMint»¹⁹. Το (TEI) ParlaMint προσφέρει ένα ισχυρό θεμέλιο για εμβάθυνση στην πολυπλοκότητα των νομοθετικών διαδικασιών, ενσωματώνοντας απρόσκοπτα δομημένα κοινοβουλευτικά δεδομένα.

Συγκριτικά, το XML LegalDocML αναπαριστά κάθε είδους επίσημο αρχείο με λεπτομερή ιεραρχική δομή. Σε αντίθεση, το TEI ParlaMint απευθύνεται ειδικά σε κοινοβουλευτικά δεδομένα, αποτυπώνοντας τις περίπλοκες νομοθετικές αποχρώσεις.

Το μοντέλο TEI ParlaMint δημιουργήθηκε ως απάντηση στην ανάγκη να ενσωματωθούν διάφορα σύνολα δεδομένων και να διασφαλιστεί η ομοιόμορφη αναπαράσταση των κοινοβουλευτικών πληροφοριών. Το ParlaMint επεκτείνει την μορφή του TEI ώστε να προσαρμόζεται στις ιδιαιτερότητες των νομοθετικών δεδομένων, δημιουργώντας μια ενιαία πλατφόρμα για μελέτη και σύγκριση.

Πέρα από την εμφανή διαλειτουργικότητα, το ParlaMint έχει την άνεση να συλλέγει δεδομένα σχετικά με τη νομοθεσία σε λεπτομερές επίπεδο. Το σχήμα επιτρέπει τη λεπτομερή καταγραφή των νομοθετικών διαδικασιών, συμπεριλαμβανομένων των ομιλιών, των συζητήσεων, των τροπολογιών και των πρακτικών ψηφοφορίας. Αυτό το επίπεδο εξειδίκευσης επιτρέπει στους ερευνητές να διερευνήσουν συγκεκριμένες πτυχές των νομοθετικών ενεργειών και επιτρέπει εξελιγμένες αξιολογήσεις του περιεχομένου.

2.6 *Θεματική Μοντελοποίηση (Topic Modelling) - LDA*

Στην εποχή της πληροφορίας, η ποσότητα του κειμένου που συναντάμε καθημερινά ξεπερνά τις ικανότητες μας για επεξεργασία. Στο πλαίσιο αυτό, ο τρόπος με τον οποίο ερμηνεύονται και κατανοούνται τα δεδομένα κειμένου έχει μετασχηματιστεί πλήρως τα τελευταία χρόνια, με την ενσωμάτωση νέων μεθοδολογιών, με χαρακτηριστικό παράδειγμα την ευρέως χρησιμοποιούμενη Latent Dirichlet Allocation (LDA) για μοντελοποίηση θεμάτων.

Η μοντελοποίηση θεμάτων (topic modelling) είναι ένα εργαλείο εξόρυξης κειμένου που χρησιμοποιείται συχνά για την ανίχνευση κρυμμένων σημασιολογικών δομών σε ένα κείμενο. Δεδομένου ότι ένα κείμενο αφορά ένα συγκεκριμένο θέμα, συγκεκριμένες λέξεις

¹⁷ <https://tei-c.org/>

¹⁸ <https://www.clarin.eu/>

¹⁹ <https://www.clarin.eu/parlamint>

εμφανίζονται περισσότερο ή λιγότερο συχνά. Τα “θέματα” που παράγονται από τεχνικές μοντελοποίησης θεμάτων είναι ομάδες παρόμοιων λέξεων. Οι τεχνικές τυποποιούνται σε μαθηματικό πλαίσιο, επιτρέποντας τη εξέταση συνόλων κειμένων και την εξαγωγή θεμάτων. Τα θεματικά μοντέλα, γνωστά και ως πιθανοτικά θεματικά μοντέλα, αναφέρονται σε στατιστικούς αλγόριθμους για την ανακάλυψη των λανθανόντων σημασιολογικών δομών μέσα σε εκτεταμένα σώματα κειμένων.

Κατά την πρακτική εφαρμογή αυτών των μοντέλων, αρχικά συγκεντρώνονται μεγάλες ποσότητες δεδομένων απλού κειμένου. Συνήθως, το αδόμητο κειμενικό υλικό είναι περίπλοκο και ιδιαίτερα εκτενές. Η μη επιβλεπόμενη γεννητική πιθανοτική μέθοδος – που ονομάζεται Latent Dirichlet Allocation (LDA) – αναδεικνύεται σε βασικό εργαλείο για την επεξεργασία και οργάνωση των κειμενικών δεδομένων. Πιο συγκεκριμένα, η LDA υποθέτει ότι κάθε έγγραφο/κείμενο μπορεί να αναπαρασταθεί ως μια πιθανοτική κατανομή πάνω σε λανθάνοντα θέματα και ότι η κατανομή των θεμάτων σε όλα τα έγγραφα/κείμενα μοιράζεται μια κοινή κατανομή, γνωστή ως ‘κατανομή Dirichlet’. Κάθε λανθάνον θέμα στο μοντέλο LDA αναπαρίσταται επίσης ως μια πιθανοτική κατανομή επί λέξεων και οι κατανομές λέξεων των θεμάτων μοιράζονται και αυτή μια κοινή κατανομή Dirichlet. Με άλλα λόγια, η LDA υποθέτει ότι τα έγγραφα παράγονται χρησιμοποιώντας μια στατιστική παραγωγική διαδικασία, έτσι ώστε κάθε έγγραφο να είναι ένα μείγμα θεμάτων και κάθε θέμα να είναι ένα μείγμα λέξεων.

Η LDA εξυπηρετεί ένα πλήθος πρακτικών εφαρμογών, όπως η ομαδοποίηση εγγράφων για την ομαδοποίηση σχετικών εγγράφων με βάση κοινά θέματα, η σύνοψη θεμάτων για τη δημιουργία συνοπτικών περιλήψεων και η βελτίωση της ακρίβειας της ανάκτησης πληροφοριών με τη σύνδεση εγγράφων με σχετικά θέματα.

Βέβαια, η αποτελεσματικότητα της LDA εξαρτάται από την επιλογή των κατάλληλων υπερπαραμέτρων, κυρίως του αριθμού των θεμάτων, κάτι που επηρεάζει το αποτέλεσμα του μοντέλου. Επιπλέον, παρά την αποδοτική της λειτουργία στην ανίχνευση θεμάτων, η ανθρώπινη αλληλεπίδραση είναι απαραίτητη για την ερμηνεία αυτών των θεμάτων.

Συνοψίζοντας, η μοντελοποίηση θεμάτων μέσω της Latent Dirichlet Allocation αποτελεί απαραίτητη τεχνική στον τομέα της ανάλυσης κειμένου. Προσφέρεται ένας μεθοδικός και εξελιγμένος τρόπος για την αποκάλυψη λανθάνουσας δομής σε σύνολα δεδομένων μεγάλης κλίμακας, παρέχοντας πολύτιμες πληροφορίες για τα κειμενικά δεδομένα. Αυτό όχι μόνο υποστηρίζει την εξόρυξη γνώσης, αλλά βοηθά επίσης στη λήψη καλά τεκμηριωμένων αποφάσεων στο πλούσιο σε δεδομένα περιβάλλον μας.

3

Σύστημα Διαχείρισης Πρακτικών

Στο κεφάλαιο αυτό, επικεντρωνόμαστε στην εφαρμογή των σημασιολογικών τεχνολογιών για τη μετατροπή των συζητήσεων του Ελληνικού Κοινοβουλίου σε ανοιχτά διασυνδεδεμένα δεδομένα. Πιο συγκεκριμένα, περιγράφουμε την μεθοδολογία για την μετατροπή των κοινοβουλευτικών συζητήσεων σε δομημένα αρχεία, της μορφής xml – LegalDocML και tei – ParlaMint. Στην συνέχεια ενισχύουμε την σημασιολογική αναπαράσταση των αρχείων μας με την δόμηση τους σε αρχεία μορφής RDF αποθηκεύοντας τα έπειτα για περαιτέρω μελέτη σε ένα σύστημα που επιτρέπει την πρόσβαση στα RDF δεδομένα, το Apache Jena Fuseki. Τέλος, εξετάζουμε και τον τρόπο χρήσης των ερωτημάτων SPARQL για την εξαγωγή ορισμένων παραδειγμάτων χρήσης από τα συνδεδεμένα δεδομένα.

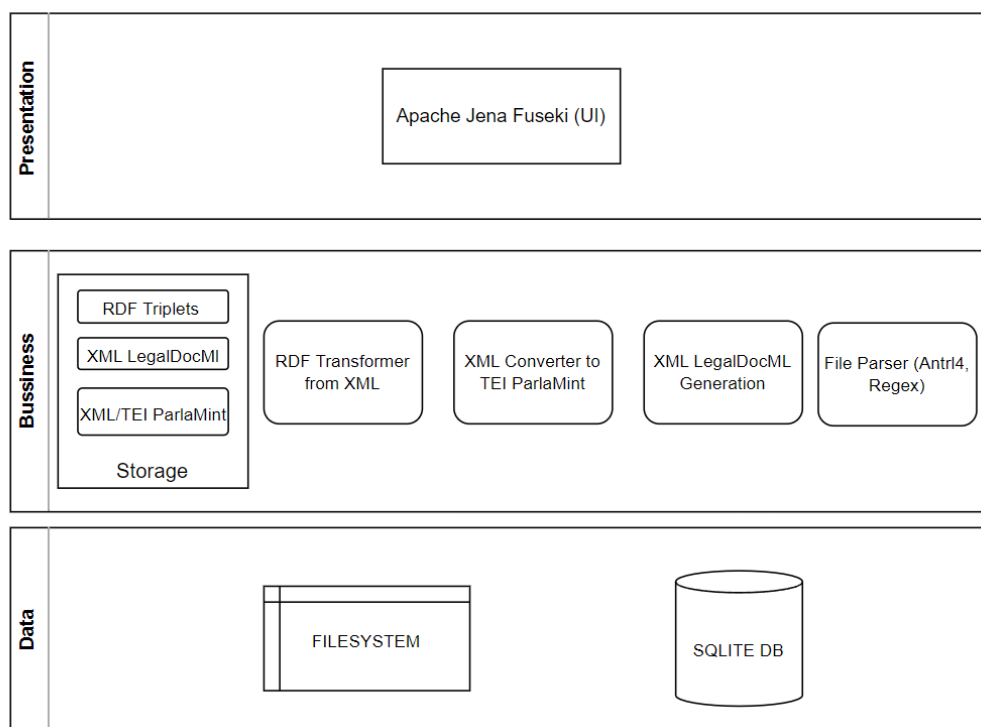
3.1 Αρχιτεκτονική Συστήματος

Η αρχιτεκτονική του συστήματος περιλαμβάνει όλα τα στοιχεία και τις τεχνολογίες που απαιτούνται για τη μετατροπή των συζητήσεων του Ελληνικού Κοινοβουλίου σε συνδεδεμένα δεδομένα. Στα διαγράμματα (εικόνες 3.1 – 3.3) παρουσιάζονται τα βασικά στοιχεία που συνθέτουν το πλαίσιο του συστήματός μας. Η προτεινόμενη αρχιτεκτονική εξυπηρετεί μια μεθοδική προσέγγιση που συνδυάζει διάφορες φάσεις και τεχνολογίες αιχμής προκειμένου να μετατρέψει τον αδόμητο κοινοβουλευτικό λόγο σε μορφή οργανωμένη και κατανοητή.

Συγκεκριμένα, στο πρώτο διάγραμμα (εικόνα 3.1) καταγράφεται η αρχιτεκτονική ολοκλήρου του συστήματος. Εμφανίζονται διακριτά ανά «γραμμή» από κάτω προς τα πάνω, οι τοποθεσίες αποθήκευσης δεδομένων, τα διάφορα υποσυστήματα και τέλος η διεπαφή του χρήστη με το σύστημα. Ως τοποθεσία αποθήκευσης στην παρούσα εργασία θεωρούμε την τοπική μνήμη του υπολογιστή, αντλώντας τα αρχεία απευθείας από τον τοπικό φάκελο. Τα αρχεία αυτά συγκεντρώνονται μέσα σε έναν φάκελο και αποτυπώνονται σε μια βάση δεδομένων SQLITE, με τη διαδρομή αποθήκευσης των αρχείων στον υπολογιστή να αναφέρεται σε πεδίο της βάσης δεδομένων.

Στην συνέχεια, στο μεσαίο τμήμα του διαγράμματος βρίσκονται τα υποσυστήματα που είναι υπεύθυνα για την συνολική λειτουργία συστήματος και την τελική παραγωγή δομημένων αρχείων. Τέλος, στην κορυφή βρίσκεται το σημείο όπου οι χρήστες επικοινωνούν με τα δομημένα κοινοβουλευτικά δεδομένα, μέσα από ένα υποσύστημα διαχείρισης και επεξεργασίας σημασιολογικών αρχείων RDF, το Apache Jena Fuseki. Το Apache Fuseki παρέχει λειτουργίες για τη αποθήκευση δεδομένων και την εκτέλεση ερωτημάτων μέσω από

ένα SPARQL endpoint που δημιουργεί. Αυτό το endpoint βρίσκεται στο βασικό UI που παράγεται αυτόματα σε συγκεκριμένη θύρα (port) κατά την εκτέλεση του Apache Jena Fuseki.

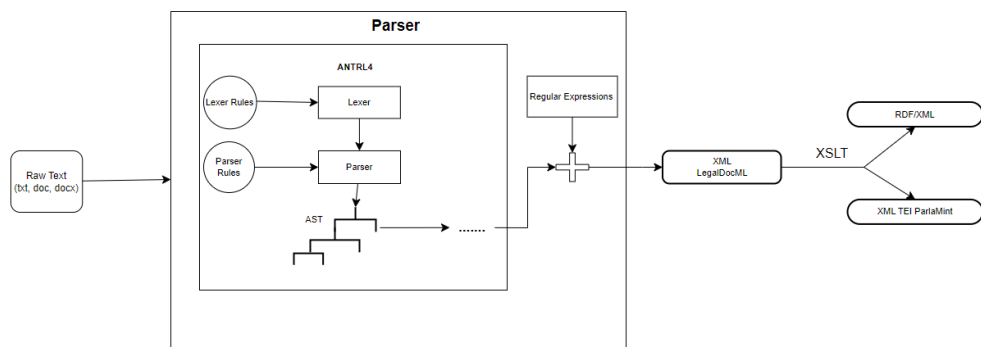


Εικόνα 3.1 – Διάγραμμα Αρχιτεκτονικής Συστήματος

Έπειτα, στο δεύτερο διάγραμμα (εικόνα 3.2) φαίνεται ένα συνολικό pipeline του συστήματος από την αρχική είσοδο των αρχείων έως και την τελική εξαγωγή όλων των μορφών αρχείων. Το πρώτο στάδιο του υποσυστήματος περιλαμβάνει την απόκτηση των συζητήσεων του Ελληνικού Κοινοβουλίου σε μορφή Word ή απλού κειμένου. Τα έγγραφα αυτά χρησιμεύουν ως δεδομένα εισόδου για τη μετέπειτα διαδικασία μετασχηματισμού. Η απόκτηση των δεδομένων μπορεί να πραγματοποιηθεί με διάφορους τρόπους, όπως η πρόσβαση σε επίσημα κοινοβουλευτικά αρχεία ή η αξιοποίηση δεδομένων που παρέχονται από εξουσιοδοτημένες πηγές, ή από ήδη υπάρχουσες εργασίες.

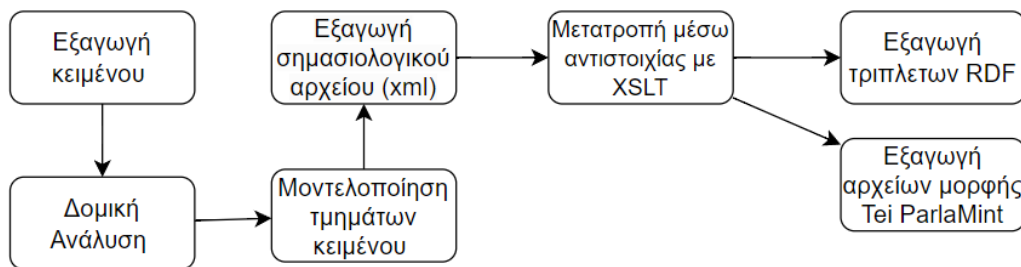
Σε δεύτερο βήμα, εξάγουμε και επεξεργαζόμαστε τα αρχικά κείμενα. Αυτά μετατρέπονται σε μια δομημένη μορφή XML, μέσω της αξιοποίησης ενός parser που κατασκευάσαμε. Αναλύουμε δομικά τα έγγραφα, και προσδιορίζοντας τα σχετικά στοιχεία και χαρακτηριστικά, παράγουμε δομημένα αρχεία XML, της μορφής LegalDocML, που αντιπροσωπεύουν τη δομημένη αναπαράσταση των συζητήσεων.

Στην συνέχεια, επεξεργαζόμαστε τα παραγόμενα αρχεία XML (LegalDocML) και τα μετασχηματίζουμε σε σημασιολογικά αρχεία που περιέχουν τριπλέτες RDF, με τον κατάλληλο συνδυασμό βιβλιοθηκών. Ο μετασχηματισμός αυτός περιλαμβάνει την αντιστοίχιση των στοιχείων και χαρακτηριστικών XML σε ιδιότητες και πόρους RDF. Τα μετασχηματισμένα δεδομένα τηρούν τη μορφή RDF Schema (RDFS), εξασφαλίζοντας τη διαλειτουργικότητα και τη συμβατότητα με τις αρχές των συνδεδεμένων δεδομένων. Παράλληλα, διενεργούμε και μετατροπή των αρχείων XML σε αρχεία της μορφής TEI ParlaMint, παρέχοντας μια νέα δομημένη μορφή των κοινοβουλευτικών συζητήσεων.



Εικόνα 3.2 – Συνολικό pipeline του συστήματος

Τέλος, στο τρίτο διάγραμμα (εικόνα 3.3), αποτυπώνεται η μεθοδολογία βημάτων που ακολουθείται για την σημασιολογική μετατροπή των κειμενικών δεδομένων σε δομημένα αρχεία. Ξεκινάμε με την δομική επεξεργασία των αρχικών κειμένων και την εξαγωγή δομημένων αρχείων, μορφής xml. Έπειτα, αξιοποιούμε τα δομημένα αρχεία, και υποβάλλοντας τα σε μετασχηματισμό μέσω αντιστοίχισης XSLT, εξάγουμε σημασιολογικά αρχεία της μορφής RDF και δομημένα αρχεία μορφής ParlaMint.



Εικόνα 3.3 – Pipeline της διαδικασίας μετασχηματισμού

3.2 Δεδομένα Συστήματος

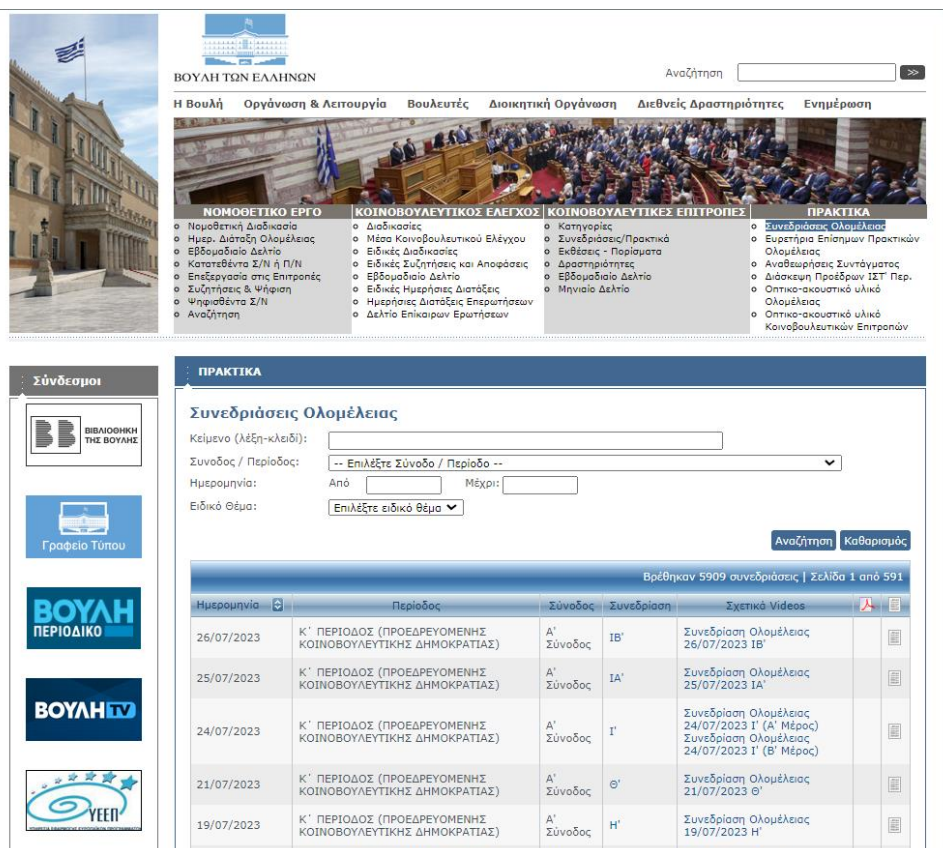
3.2.1 Βάση Δεδομένων

Η βάση δεδομένων μας περιέχει αρχεία που ανακτήθηκαν απευθείας από τον επίσημο ιστότοπο του Ελληνικού Κοινοβουλίου²⁰ (Εικόνα 3.4). Το γεγονός ότι το σύνολο των δεδομένων προέρχεται από μια επίσημη ιστοσελίδα ενισχύει τη εγκυρότητα των δεδομένων και επιτρέπει την σφαιρική κατανόησή του συνεχώς μεταβαλλόμενου πολιτικού περιβάλλοντος της Ελλάδας. Ο ιστότοπος παρέχει αρχεία από το 1989 και συνεχίζει να αποτελεί πολύτιμη πηγή για την κατανόηση των πολιτικών εξελίξεων στην Ελλάδα μέχρι και σήμερα. Για τους

²⁰ <https://www.hellenicparliament.gr/Praktika/Synedriaseis-Olomeleias>

σκοπούς της παρούσας εργασίας, συμπεριλάβαμε αρχεία από το 1989 έως και τις αρχές Νοεμβρίου 2023, καλύπτοντας ένα διάστημα άνω των τριών δεκαετιών.

Ο συνολικός αριθμός των αρχείων του συνόλου δεδομένων μας είναι περίπου 5.800, τα οποία είναι διαθέσιμα σε μορφή doc(x) και txt. Να σημειώσουμε ότι το σύνολο δεδομένων περιλαμβάνει αρχεία από όλα έτη, με εξαίρεση το έτος 1995, για το οποίο δεν υπάρχουν διαθέσιμα αρχεία στον επίσημο ιστότοπο.



Εικόνα 3.4 – Επίσημος ιστότοπος Ελληνικής Κυβέρνησης με Συνεδριάσεις Ολομέλειας

3.2.2 Δομή Πρακτικών Βουλής

Η δομή των Πρακτικών του Ελληνικού Κοινοβουλίου ακολουθεί μια καλά οργανωμένη και ιεραρχική μορφή, παρέχοντας μια ολοκληρωμένη επισκόπηση των νομοθετικών συζητήσεων και αντιπαραθέσεων που λαμβάνουν χώρα στο Κοινοβούλιο. Επειδή όμως το διάστημα μελέτης είναι μεγάλο (1989 μέχρι και αρχές Νοέμβριου του 2023) παρατηρούμε μια εναλλαγή στον τρόπο συγγραφής τους. Για αυτό το λόγο η ανάλυση δεν μπορεί να είναι πλήρης, αλλά καλύπτει ένα μεγάλο μέρος των αρχείων, μιας και κάνουμε εκτεταμένη ανάλυση με αρκετές από τις υπο-κατηγορίες που εντοπίστηκαν.

Στην αρχή, συναντάμε τον «ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ» (Εικόνα 3.5) που περιλαμβάνει συνήθως βασικές πληροφορίες, όπως η περίοδος που διαδραματίζεται η συνεδρίαση (π.χ. ΙΖ' ΠΕΡΙΟΔΟΣ), ο τύπος της κυβέρνησης (π.χ. ΠΡΟΕΔΡΙΚΗ ΚΟΙΝΟΒΟΥΛΕΥΤΙΚΗ ΔΗΜΟΚΡΑΤΙΑ), η σύννοδος (π.χ. Σύννοδος Γ') και ο αριθμός της

συνεδρίασης (π.χ. ΣΥΝΕΔΡΙΑΣΗ ΡΓΛ'). Επιπλέον, περιέχει τη συγκεκριμένη ημερομηνία κατά την οποία έλαβε χώρα η κοινοβουλευτική σύνοδος, όπως "Παρασκευή 8 Ιουνίου 2018". Η ημερομηνία αυτή χρησιμεύει ως κρίσιμο σημείο αναφοράς για τους αναγνώστες προκειμένου να εντοπίσουν και να αποκτήσουν πρόσβαση στις συζητήσεις που διεξήχθησαν κατά τη διάρκεια της συγκεκριμένης ημερομηνίας.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ
ΙΖ' ΠΕΡΙΟΔΟΣ
ΠΡΟΕΔΡΕΥΟΜΕΝΗΣ ΚΟΙΝΟΒΟΥΛΕΥΤΙΚΗΣ ΔΗΜΟΚΡΑΤΙΑΣ
ΣΥΝΟΔΟΣ Γ'

ΣΥΝΕΔΡΙΑΣΗ ΡΛΓ'
Παρασκευή 8 Ιουνίου 2018

Εικόνα 3.5 – Ενδεικτικό απόσπασμα πρακτικών – «ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ» (08-06-2018)

Στις επόμενες γραμμές, καταγράφεται, συνήθως, μια λίστα με τα θέματα που συζητήθηκαν κατά τη διάρκεια της συνεδρίασης, κατηγοριοποιημένα με βάση τη συνάφεια και το αντικείμενό τους (Εικόνα 3.6). Κάθε θέμα μπορεί να περιλαμβάνει μια σύντομη περιγραφή, παρέχοντας στους αναγνώστες μια επισκόπηση των θεμάτων που εξετάστηκαν, δίνοντας μια πρώτη ματιά για το τι συζητήθηκε στην συνεδρίαση.

ΘΕΜΑΤΑ

A. ΕΙΔΙΚΑ ΘΕΜΑΤΑ

1. Επικύρωση Πρακτικών, σελ.
2. Άδεια απουσίας του Βουλευτή κ. Θ. Θεοχάρη, σελ.
3. Ανακοινώνεται ότι τη συνεδρίαση παρακολουθούν μαθητές από το 89ο Δημοτικό Σχολείο Αθήνας, το Δημοτικό Σχολείο Γαλιός Ηρακλείου, το 3ο Δημοτικό Σχολείο Χίου και το 1ο Δημοτικό Σχολείο Καρύστου Ευβοίας, σελ.
4. Κατάθεση από τον κ. Κ. Σκανδαλίδη, επιστολής του κ. Γιάννη Παπακωνσταντίνου - Γενικού Διευθυντή του ΠΑΣΟΚ κατά το έτος 2007 - σχετικά με την δημοσίευση αθωωτικής απόφασης του Εφετείου Αθηνών, σελ.
5. Αναφορά στην επίθεση στο γραφείο του Βουλευτή κ. Μ. Βαρβιτσιώτη και καταδίκη αυτής, σελ.
6. Επί διαδικαστικού θέματος, σελ.

B. ΚΟΙΝΟΒΟΥΛΕΥΤΙΚΟΣ ΕΛΕΓΧΟΣ

1. Ανακοίνωση αναφορών, σελ.
2. Συζήτηση επίκαιρης ερώτησης προς τον Υπουργό Ψηφιακής Πολιτικής, Τηλεπικοινωνιών και Ενημέρωσης, με θέμα: «Ταλαιπωρία και επιβάρυνση των καταναλωτών από την καθυστέρηση απόδοσης χρηματικών ποσών που έχουν καταβάλει στα ΕΛΤΑ για εξόφληση λογαριασμών της ΔΕΗ», σελ.
3. Συζήτηση της υπ' αριθμόν 26/17/24-5-2018 επίκαιρης επερώτησης, που κατέθεσαν ο Πρόεδρος της κοινοβουλευτικής ομάδας και Γραμματέας της

Εικόνα 3.6 – Ενδεικτικό απόσπασμα πρακτικών – «Θέματα» (08-06-2018)

Μετά τον κατάλογο θεμάτων, στα πρακτικά παρουσιάζεται μια πλήρη λίστα όλων των ατόμων που διετέλεσαν πρόεδροι/προεδρεύοντες στην συνεδρίαση. Έπειτα, καταγράφονται τα ονόματα όλων των ομιλητών που συμμετείχαν στις συζητήσεις (Εικόνα 3.7). Ο κατάλογος αυτός επιτρέπει στους αναγνώστες να εντοπίσουν τους βασικούς συμμετέχοντες.

ΠΡΟΕΔΡΕΥΟΝΤΕΣ

ΚΑΜΜΕΝΟΣ Δ. , σελ.
ΚΡΕΜΑΣΤΙΝΟΣ Δ. , σελ.

ΟΜΙΛΗΤΕΣ

Α. Επί της αναφοράς στην επίθεση στο γραφείο του Βουλευτή κ. Μ.

Βαρβιτσιώτη

ΚΑΜΜΕΝΟΣ Δ. , σελ.

ΛΟΒΕΡΔΟΣ Α. , σελ.

ΛΥΚΟΥΔΗΣ Σ. , σελ.

Β. Επί διαδικαστικού θέματος:

ΚΑΜΜΕΝΟΣ Δ. , σελ.

ΚΡΕΜΑΣΤΙΝΟΣ Δ. , σελ.

Γ. Επί της επίκαιρης ερώτησης:

ΑΣΗΜΑΚΟΠΟΥΛΟΥ Α. , σελ.

Εικόνα 3.7 – Ενδεικτικό απόσπασμα πρακτικών – «Προεδρεύοντες» και «Ομιλητές» (08-06-2018)

Στην συνέχεια, συνήθως υπάρχει μια υπο-ενότητα «ΠΡΑΚΤΙΚΑ ΒΟΥΛΗΣ», όπου καταγράφονται οι ίδιες πληροφορίες που αναφέρονται στον «ΠΙΝΑΚΑ ΠΕΡΙΕΧΟΜΕΝΩΝ». Αναφέρεται εκ νέου την περίοδο, τον τύπο της κυβέρνησης, τη σύνοδο, τον αριθμό της συνεδρίασης και την ημερομηνία για λόγους αναφοράς.

Έπειτα, υπάρχει ένας σύντομος πρόλογος (Εικόνα 3.8), όπου ακολουθεί πάντα την ίδια αυστηρή και τυπική μορφή, στην οποία παρέχονται βασικές πληροφορίες για την συνεδρίαση. Ξεκινάει με τον τόπο, την ημερομηνία και την ώρα, και αφού έχει προλογηθεί η συνεδρίαση ολοκληρώνεται η παράγραφος με την παρουσίαση του προέδρου, όπου για παράδειγμα στο παρόν είναι ο κ. «Δημήτριος Κρεμαστινός».

Αθήνα, σήμερα στις 8 Ιουνίου 2018, ημέρα Παρασκευή και ώρα 10.20΄,

συνήλθε στην Αίθουσα των συνεδριάσεων του Βουλευτηρίου η Βουλή σε

ολομέλεια για να συνεδριάσει υπό την προεδρία του Ε΄ Αντιπροέδρου αυτής κ.

ΔΗΜΗΤΡΙΟΥ ΚΡΕΜΑΣΤΙΝΟΥ.

Εικόνα 3.8 – Ενδεικτικό απόσπασμα πρακτικών – Εισαγωγικός Πρόλογος (08-06-2018)

Έπειτα και αφότου έχουν ολοκληρωθεί τα βασικά συστατικά της εισαγωγής, ακολουθούν οι διάλογοι που αποτυπώνουν τις ομιλίες, τις παρεμβάσεις και τις συζητήσεις των διαφόρων βουλευτών που συμμετείχαν στη σύνοδο. Πάντα τον πρώτο λόγο έχει ο προεδρεύων που ορίστηκε στην προηγούμενη εισαγωγική παράγραφο. Είναι το άτομο που οριοθετεί την

συζήτηση ξεκινώντας την παρουσιάζοντας κάποιες φορές τις επίκαιρες ερωτήσεις και την ολοκληρώνει με την λήξη της συνεδρίασης.

Στο ενδιαμέσο και κύριο πυλώνα των πρακτικών, όπου είναι αυτό που περιέχονται οι διάλογοι χρησιμοποιείται συνήθως μια ιδιαίτερη σύμβαση μορφοποίησης για την επισήμανση των ονομάτων των ομιλητών. Το όνομα κάθε ομιλητή καταγράφεται με έντονη μαύρη γραφή, εξασφαλίζοντας την αναγνώρισή του μέσα στο κείμενο. Αυτή η οπτική έμφαση εξυπηρετεί την εύκολη διάκριση των ομιλητών και επιτρέπει στους αναγνώστες να παρακολουθούν αποτελεσματικότερα τη ροή των συζητήσεων. Κάποιες φορές περιλαμβάνονται πρόσθετες πληροφορίες για τους ομιλητές, οι οποίες παρουσιάζονται εντός παρενθέσεων έπειτα από το όνομα. Αυτή η συμπληρωματική πληροφορία μπορεί να είναι ένα σχόλιο που δηλώνει την κοινοβουλευτική θέση ή την ιδιότητα του ομιλητή. Η μόνη διαφορά είναι στον πρόεδρο, μιας και συμβαίνει το ανάποδο, όπου για λόγους έμφασης δίνεται πρώτα η προεδρική ιδιότητα. Αυτές οι συμπληρωματικές πληροφορίες παρέχουν πολύτιμο πλαίσιο για την κατανόηση του ρόλου και του κύρους κάθε συμμετέχοντα στις κοινοβουλευτικές συζητήσεις.

Ένα παράδειγμα βρίσκεται στη εικόνα 3.9, όπου φαίνεται ένα απόσπασμα διάλογου τριών ατόμων, ανάμεσα στους «Άννα-Μισέλ Ασημακοπούλου», «Δημήτριος Κρεμαστινός» και «Νικόλαος Παππάς». Γίνεται εμφανές η μορφοποίηση που αναλύθηκε προηγουμένως, καθώς τα ονόματα και οι ιδιότητες/αξιώματα αποτυπώνονται σε έντονη χαρακτηριστική μορφή.

ANNA - ΜΙΣΕΛ ΑΣΗΜΑΚΟΠΟΥΛΟΥ: Αυτά έχουν καταβληθεί. Πολύ
ωραία. Δέχομαι την απάντηση, αλλά θα ήθελα να ξέρω το εξής: Υπάρχουν
ενέργειες για το μέλλον; Υπάρχει κάτι το οποίο θα μας εξασφαλίσει ότι ο κόσμος
δεν θα ξαναβρίσκεται εκεί; Ό,τι και να συζητάμε, όσο και να συγκρούμαστε
εδώ μέσα, αυτός που κάθεσαι στην ουρά θέλει να έχει μια απάντηση για το τι
θα γίνει όταν θα πάει να ~~ξεναποληρώσει~~ τον λογαριασμό του στον ΕΛΤΑ.

Ευχαριστώ.

ΠΡΟΕΔΡΕΥΩΝ (Δημήτριος Κρεμαστινός): Ευχαριστούμε.

Κύριε Υπουργέ, έχετε τον λόγο.

**ΝΙΚΟΛΑΟΣ ΠΑΠΠΑΣ (Υπουργός Ψηφιακής Πολιτικής,
Τηλεπικοινωνιών και Ενημέρωσης):** Ευχαριστώ, κύριε Πρόεδρε.

Κυρία Ασημακοπούλου, νομίζω ότι δεν με παρακολουθήσατε.

Εικόνα 3.9 – Ενδεικτικό απόσπασμα πρακτικών – Μέρος Διαλόγων (08-06-2018)

3.3 Διαχείριση και Μετατροπή Δεδομένων

Στην ενότητα αυτή παρουσιάζουμε τα στάδια της εξαγωγής του κειμένου, και έπειτα συνεχίζουμε με την ανάλυση για την διαχείριση και μετατροπή των αρχείων σε ανοιχτά διασυνδεδεμένα δεδομένα.

3.3.1 Εξαγωγή και Ανάλυση Κειμένων

Σε πρώτο στάδιο δημιουργούμε έναν υποσύστημα – με έναν *parser* – με σκοπό την ανάλυση και την επεξεργασία των πρακτικών του Ελληνικού Κοινοβουλίου. Για αρχή, αναλύουμε τα ακατέργαστα κειμενικά αρχεία που βρίσκονται στον αρχικό φάκελο με την βιβλιοθήκη *tika-python*, μια έκδοση της Apache Tika²¹ για γλώσσα Python. Έπειτα, τα επεξεργαζόμαστε με σκοπό να διαχωρίσουμε το εκάστοτε έγγραφο και να εξάγουμε τμήματά του. Για αυτό το βήμα χρησιμοποιούμε την γεννήτρια ανάλυσης ANTLR-4 και τις κανονικές εκφράσεις (regular expressions).

Τα εισαγωγικά τμήματα διακρίνονται από το κύριο σώμα των πρακτικών στην αρχή της διαδικασίας ανάλυσης, με την γραμματική που έχουμε συντάξει. Η σύνοδος, η ημερομηνία και ο αριθμός της συνεδρίασης είναι μεταξύ των κρίσιμων μεταδεδομένων που αναγνωρίζουμε και ανακτούμε σε πρώτο στάδιο.

Στην συνέχεια, ομαδοποιούμε με ακρίβεια κάθε ομιλητή με τον αντίστοιχο λόγο του αξιοποιώντας τις λειτουργίες των «Κανονικών Εκφράσεων» (Regular expression). Πιο συγκεκριμένα, με αυτοματοποιημένο τρόπο αναγνωρίζουμε κάθε φορά το όνομα του ομιλητή, μαζί με την ιδιότητα του αν έχει, ακολουθούμενο από μια άνω και κάτω τελεία (':'). Στη συνέχεια καταγράφουμε ολόκληρη την ομιλία του, καθορίζοντας έτσι σαφώς τα όρια της συμμετοχής του κάθε ατόμου στο πλαίσιο των κοινοβουλευτικών διαδικασιών.

Η αξιοποίηση των regex εκφράσεων επικεντρώνεται κυρίως στην αναγνώριση του ομιλητή, του προέδρου, του λόγου ή σχόλιων όπως χειροκροτημάτων, ηλεκτρονικής καταμέτρησης. Σε όποιο σημείο χρησιμοποιούμε για την αναγνώριση κανονικές εκφράσεις, είναι λέξεις-φράσεις όπου ακολουθούν μια συγκεκριμένη δομή ή μορφοποίηση στην έκφραση τους. Έτσι, η τυποποίηση τους είναι δεδομένη, μιας και επιτυγχάνεται η αντιστοίχιση συγκεκριμένων μοτίβων.

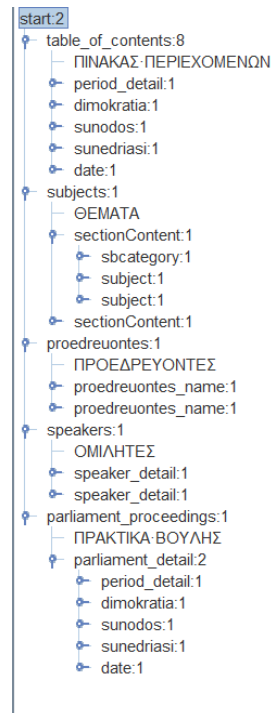
Επομένως φαίνεται πως η αρχική διαδικασία ανάλυσης που δημιουργήσαμε με τον parser, αναδεικνύεται αναγκαία στην διαδικασία αυτοματοποίησης της μετατροπής των νομοθετικών αρχείων σε συνδεδεμένα δεδομένα. Καταφέραμε να διαχειριστούμε αποτελεσματικά τις κοινοβουλευτικές διαδικασίες, διατηρώντας την ακρίβεια των εξαγόμενων δεδομένων και συμβάλλοντας στην παροχή μιας συνδεδεμένης απεικόνισης των συζητήσεων στο Ελληνικό Κοινοβούλιο.

3.3.1.1 Γλώσσα Περιγραφής Κοινοβουλευτικών Αρχείων

Για να εντοπίσουμε τις εισαγωγικές λεπτομέρειες μέσα στα κοινοβουλευτικά έγγραφα, αναπτύξαμε μια γλώσσα ειδικού σκοπού για την αναγνώριση συγκεκριμένων τμημάτων των συζητήσεων. Η γλώσσα αυτή αποτελείται από ένα σύνολο κανόνων γραμματικής που βασίζονται στις λειτουργίες της ANTLR. Η ANTLR (ANother Tool for Language Recognition), είναι ένα εργαλείο το οποίο χρησιμοποιείται κυρίως για τη δημιουργία αναλυτών για διάφορες γλώσσες προγραμματισμού και γλώσσες ειδικού τομέα (DSLs). Η προσαρμοστικότητα της επιτρέπει τη διαφοροποιημένη αναγνώριση, εξασφαλίζοντας ακριβή αναγνώριση και κατηγοριοποίηση.

²¹ <https://tika.apache.org/>

Τα βασικά αρχεία τα οποία πρέπει να δημιουργηθούν για την ολοκληρωμένη λειτουργία αυτής της γλώσσας είναι μία πλήρη γραμματική για την ANTLR4, η οποία απαρτίζεται από ένα αρχείο parser και ένα αρχείο lexer. Ο lexer σαρώνει την είσοδο και την αναλύει σε μια ροή από ‘tokens’, ενώ ο parser επεξεργάζεται αυτά τα tokens σύμφωνα με τους κανόνες της γραμματικής για να δημιουργήσει ένα αφηρημένο συντακτικό δέντρο (AST). Ένα ενδεικτικό δέντρο που προκύπτει από την γραμματική για την παρούσα εργασία είναι το παρακάτω (Εικόνα 3.10), το οποίο περιέχει αρκετές από τις υποκατηγορίες που δύνανται να αναγνωριστούν. Φαίνεται πως κάθε στοιχείο το έχουμε αναγνωρίσει και αντιστοιχίσει στην συγκεκριμένη ετικέτα-μεταβλητή.



Εικόνα 3.10 – Ενδεικτικό δέντρο που προκύπτει από την γραμματική

```
lexer grammar DebateGrammarLexer;
WS: [- \t\r\n]+ -> skip;

SIMIOSI: SPACES ('(Σημείωση:' | '(ΣΗΜΕΙΩΣΗ:') SPACES ANY_TEXT;
PINAKAS_PERIEXOMENON:
    SPACES ('ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ' | 'ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ') SPACES;

// ----- THEMATA
THEMATA_SPACES: 'ΘΕΜΑΤΑ' SPACES -> pushMode(subjects);

// ----- SPEAKERS
OMILITES: (SPACES 'ΟΜΙΛΗΤΕΣ' SPACES);
PROEDREUONTES: (
    (
        'ΠΡΟΕΔΡΕΥΟΝΤΕΣ'
        | 'ΠΡΟΕΔΡΕΥΟΝΤΕΣ'
        | 'ΠΡΟΕΔΡΕΥΩΝ'
        | 'ΠΡΟΕΔΡΕΥΟΥΣΑ'
        | 'ΠΡΟΕΔΡΟΥΣΑ'
        | 'ΠΡΟΕΔΡΟΣ'
    ) SPACES
);
```

Εικόνα 3.11 – Απόσπασμα από το συντακτικό της γραμματικής

Στο παραπάνω απόσπασμα (εικόνα 3.11), παρατίθεται ένα μέρος κανόνων του lexer για το συντακτικού της γραμματικής μας σε μορφή *Backus–Naur form*. Ενδεικτικά, φαίνονται οι κανόνες για την σημείωση, τον πίνακα περιεχομένων, για τα θέματα, για τους ομιλητές και για τους προεδρεύοντες.

3.3.2 Μετατροπή Αρχείων Κειμένου σε XML αρχεία

Σε πρώτο βήμα, καλούμαστε να μετατρέψουμε τις αδόμετες συζητήσεις του Ελληνικού Κοινοβουλίου σε δομημένες μορφές και συγκεκριμένα σε XML. Επιλέξαμε αυτά τα αρχεία να ακολουθούν τις απαιτήσεις του LegalDocML. Με βάση την αρχική διαδικασία ανάλυσης, όπως σκιαγραφήθηκε παραπάνω, έχουμε εξάγει βασικά δεδομένα από τις κοινοβουλευτικές συνεδριάσεις, συμπεριλαμβανομένων των εισαγωγικών πληροφοριών, των ονομάτων των ομιλητών και του περιεχομένου των ομιλιών τους. Αυτές οι μεταβλητές που περισυλλέξαμε αποτελούν τη βάση για τη δημιουργία οργανωμένων αρχείων XML, που συμμορφώνονται με τα πρότυπα του LegalDocML.

Τα παραγόμενα δομημένα αρχεία εξασφαλίζουν μια ομοιόμορφη αναπαράσταση των νομοθετικών συζητήσεων με τη μεθοδική δόμηση XML σύμφωνα με τις κατευθυντήριες γραμμές LegalDocML, όπως χαρακτηριστικά φαίνονται στην εικόνα 3.14 όπου δείχνει την ιεραρχική δομή που πρέπει να έχει κάθε αρχείο – debate. Όλα τα δεδομένα των αρχικών αρχείων αντιστοιχίζονται στα κατάλληλα στοιχεία XML.

Πιο συγκεκριμένα, το μέρος των εισαγωγικών πληροφοριών, όπως ημερομηνία, αριθμός συνόδου κ.α. χρησιμοποιούνται σαν μεταδεδομένα για τα LegalDocML αρχεία και τοποθετούνται κατάλληλα εντός της ετικέτας “<meta>”. Ενδεικτικά, να αναφέρουμε πως μέρος των μεταδεδομένων αποτελεί η λίστα με όλους του ομιλητές της εκάστοτε συνεδρίασης, οι οποίοι προστίθεται σε μία εξιδεικευμένη ετικέτα του προτύπου LegalDocML, την “<TLCPerson>”

```
<akoma:ttoso xmlns="http://docs.oasis-open.org/legaldocml/ns/akn/3.0">
  <debate name="debate">
    <meta>
      <identification source="#cobalt">
        <FRBRWork>
          <FRBRthis value="/akn/gr/debate/2018-06-08/1/main"/>
          <FRBRuri value="/akn/gr/debate/2018-06-08/1"/>
          <FRBRalias value="ΠΡΑΚΤΙΚΑ ΒΟΥΛΗΣ 2018-06-08" name="title"/>
          <FRBRdate date="2018-06-08" name="Generation"/>
          <FRBRauthor href="" />
          <FRBRcountry value="gr"/>
          <FRBRnumber value="1"/>
        </FRBRWork>
        <FRBRExpression>
          <FRBRthis value="/akn/gr/debate/2018-06-08/1/gr@2018-06-08/main"/>
          <FRBRuri value="/akn/gr/debate/2018-06-08/1/gr@2018-06-08"/>
          <FRBRdate date="2018-06-08" name="Generation"/>
          <FRBRauthor href="" />
          <FRBRlanguage language="gr"/>
        </FRBRExpression>
        <FRBRManifestation>
          <FRBRthis value="/akn/gr/debate/2018-06-08/1/gr@2018-06-08/main"/>
          <FRBRuri value="/akn/gr/debate/2018-06-08/1/gr@2018-06-08"/>
          <FRBRdate date="2018-06-08" name="Generation"/>
          <FRBRauthor href="" />
        </FRBRManifestation>
      </identification>
      <references source="#cobalt">
        <TLCOrganization eId="hellenic parliament" href="https://www.hellenicparliament.gr/en/" showAs="hellenic parliament"/>
        <TLCPerson eId="ioannis broutsis" href="https://www.wikidata.org/wiki/Q12875212" showAs="ΙΩΑΝΝΗΣ ΒΡΟΥΤΣΗΣ"/>
        <TLCPerson eId="anna-misel asimakopoulou" href="https://www.wikidata.org/wiki/Q16329215" showAs="ΑΝΝΑ-ΜΙΣΕΛ ΑΣΗΜΑΚΟΠΟΥΛΟΥ"/>
        <TLCPerson eId="athanasios bardalis" href="https://www.wikidata.org/wiki/Q20995539" showAs="ΑΘΑΝΑΣΙΟΣ ΒΑΡΔΑΛΗΣ"/>
        <TLCPerson eId="grigorios stogiannidis" href="https://www.wikidata.org/wiki/Q44056664" showAs="ΓΡΗΓΟΡΙΟΣ ΣΤΟΓΙΑΝΝΙΔΗΣ"/>
        <TLCPerson eId="athanasios illopoulos" href="" showAs="ΑΘΑΝΑΣΙΟΣ ΗΛΙΟΠΟΥΛΟΣ"/>
        <TLCPerson eId="ioannis delis" href="https://www.wikidata.org/wiki/Q20836555" showAs="ΙΩΑΝΝΗΣ ΔΕΛΗΣ"/>
        <TLCPerson eId="georgios lamproulis" href="https://www.wikidata.org/wiki/Q21031017" showAs="ΓΕΩΡΓΙΟΣ ΛΑΜΠΡΟΥΛΗΣ"/>
        <TLCPerson eId="iliass panagiotaros" href="https://www.wikidata.org/wiki/Q12895227" showAs="ΗΛΙΑΣ ΠΑΝΑΓΙΩΤΑΡΟΣ"/>
        <TLCPerson eId="efi axtsioglou" href="https://www.wikidata.org/wiki/Q27733766" showAs="ΕΦΗ ΑΧΤΣΙΟΓΛΟΥ"/>
        <TLCPerson eId="spuridon lukoudis" href="https://www.wikidata.org/wiki/Q18396353" showAs="ΣΠΥΡΙΔΩΝ ΛΥΚΟΥΔΗΣ"/>
        <TLCPerson eId="xristos katsotis" href="https://www.wikidata.org/wiki/Q20127736" showAs="ΧΡΗΣΤΟΣ ΚΑΤΣΙΩΤΗΣ"/>
        <TLCPerson eId="nikolaos karathanasopoulos" href="https://www.wikidata.org/wiki/Q20647546" showAs="ΝΙΚΟΛΑΟΣ ΚΑΡΑΘΑΝΑΣΟΠΟΥΛΟΣ"/>
        <TLCPerson eId="andreas loberdos" href="https://www.wikidata.org/wiki/Q499341" showAs="ΑΝΔΡΕΑΣ ΛΟΒΕΡΔΟΣ"/>
        <TLCPerson eId="ioi oi bouleutes" href="" showAs="ΟΙ ΟΙ ΒΟΥΛΕΥΤΕΣ"/>
        <TLCPerson eId="dimitrios kreastinos" href="https://www.wikidata.org/wiki/Q19635572" showAs="ΔΗΜΗΤΡΙΟΣ ΚΡΕΜΑΣΤΙΝΟΣ"/>
        <TLCPerson eId="dimitrios kammenos" href="https://www.wikidata.org/wiki/Q31284012" showAs="ΔΗΜΗΤΡΙΟΣ ΚΑΜΜΕΝΟΣ"/>
        <TLCPerson eId="nikolaos pappas" href="https://www.wikidata.org/wiki/Q2036076" showAs="ΝΙΚΟΛΑΟΣ ΠΑΠΑΪΣ"/>
        <TLCPerson eId="konstantinos skandalidis" href="https://www.wikidata.org/wiki/Q2337038" showAs="ΚΩΝΣΤΑΝΤΙΝΟΣ ΣΚΑΝΔΑΛΙΔΗΣ"/>
        <TLCPerson eId="emmanouil suntoukakis" href="https://www.wikidata.org/wiki/Q1696884" showAs="ΕΜΜΑΝΟΥΗΛ ΣΥΝΤΟΥΚΑΚΗΣ"/>
        <TLCPerson eId="ioannis gkiokas" href="https://www.wikidata.org/wiki/Q20890885" showAs="ΙΩΑΝΝΗΣ ΓΚΙΟΚΑΣ"/>
        <TLCPerson eId="dimitrios koutsoumpas" href="https://www.wikidata.org/wiki/Q10928455" showAs="ΔΗΜΗΤΡΙΟΣ ΚΟΥΤΣΟΥΜΠΑΣ"/>
      </references>
    </meta>
  </debate>
</akoma:ttoso>
```

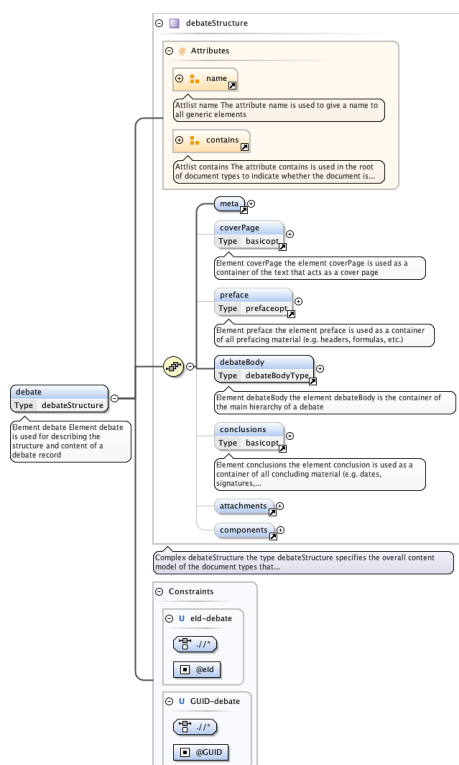
Εικόνα 3.12 – Απόσπασμα από τα μεταδεδομένα ενός αρχείου LegalDocML

Παρακάτω, (Εικόνα 3.13), φαίνεται το τελικό αποτέλεσμα ενός δομημένου αρχείου xml για το κύριο μέρος του πρακτικού μιας συνεδρίασης. Αποτελεί συνέχεια της εικόνας 3.9, στην οποία φαίνονται οι ίδιοι διάλογοι ανάμεσα στους τρεις ομιλητές.

```
<speech by="anna-misel_asimakopoulou" eId="debate_2018-06-08_1_speech_12">
  <from>ANNA-ΜΙΣΕΛ ΑΣΗΜΑΚΟΠΟΥΛΟΥ</from>
  <p>Αυτά έχουν καταβληθεί. Πολύ ωραία. Δέχομαι την απάντηση, αλλά θα ήθελα να ξέρω
  το εξής: Υπάρχουν ενέργειες για το μέλλον; Υπάρχει κάτι το οποίο θα μας εξασφαλίσει
  ότι ο κόσμος δεν θα ξαναβρίσκεται εκεί; Ό,τι και να συζητάμε, όσο και να
  συγκρουόμαστε εδώ μέσα, αυτός που κάθεται στην ουρά θέλει να έχει μια απάντηση για
  το τι θα γίνει όταν θα πάει να ξαναπληρώσει τον λογαριασμό του στον ΕΛΤΑ.</p>
  <p>Ευχαριστώ.</p>
</speech>
<speech by="dimitrios_kremastinos" eId="debate_2018-06-08_1_speech_13">
  <from>Δημήτριος Κρεμαστίνος</from>
  <p>Ευχαριστούμε.</p>
  <p>Κύριε Υπουργέ, έχετε τον λόγο.</p>
</speech>
<speech by="nikolaos_pappas" eId="debate_2018-06-08_1_speech_14">
  <from>ΝΙΚΟΛΑΟΣ ΠΑΠΠΑΣ</from>
  <p>Ευχαριστώ, κύριε Πρόεδρε.</p>
  <p>Κυρία Ασημακοπούλου, νομίζω ότι δεν με παρακολουθήσατε.</p>
</speech>
```

Εικόνα 3.13 – Απόσπασμα από κύριο μέρος (debateBody) ενός αρχείου LegalDocML

Σε αυτό το σημείο και όσον αφορά την τελευταία εικόνα (εικόνα 3.13), κάθε λόγος αποδίδεται σε ομιλητή μέσω του “by” και επισημαίνεται με ένα αναγνωριστικό “eId” όπου ακολουθεί την σύνταξη *eId* = *debate_YYYY-MM-DD_debateNum_speech_id*, με *debateNum* να είναι ο αύξων αριθμός εγγράφου εκείνη την ημέρα και *id* να είναι ο αύξων αριθμός του speech στο κάθε debate xml αρχείο. Έπειτα στην συνέχεια, βρίσκεται η ετικέτα “<from>” στην οποία υπάρχει το όνομα του ομιλητή αυτούσιο όπως στο κείμενο και οι υπόλοιπες γραμμές είναι αφιερωμένες στο κύριο μέρος του λόγου, μέσα σε ετικέτες “<p>”.



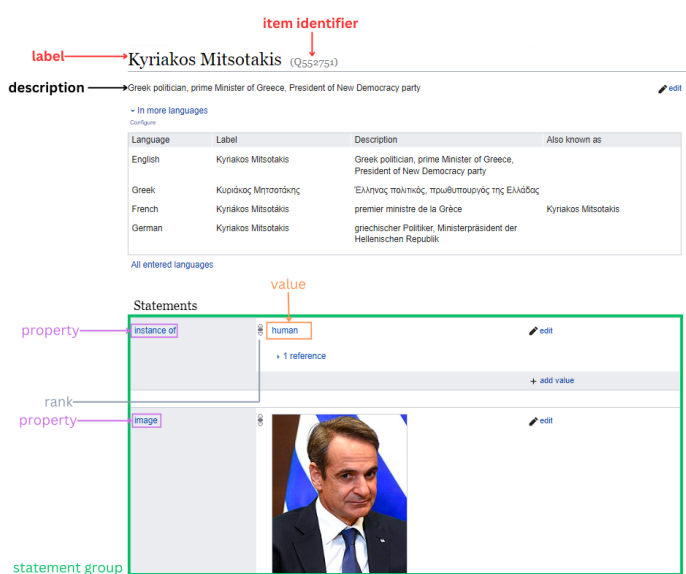
Εικόνα 3.14 – Διάγραμμα για το στοιχείο “debate”²²

²²Πηγή: https://docs.oasis-open.org/legaldocml/akn-core/v1.0/os/part2-specs/os-part2-specs_xsd_Element_debate.html

3.3.2.1 Διασύνδεση με Wikidata

Στον τομέα της αναπαράστασης γνώσης και της διασύνδεσης δεδομένων, η σύνδεση δεδομένων με πληροφορίες διαφορετικών πηγών είναι καίριας σημασίας. Μία τέτοια εναλλακτική πηγή είναι το “Wikidata”, μια ελεύθερη και ανοικτή βάση γνώσης που μπορεί να διαβαστεί και να επεξεργαστεί τόσο από ανθρώπους όσο και από μηχανές. Με άλλα λόγια, είναι μια συλλογή άρθρων που αποθηκεύονται σε μια βάση δεδομένων προσανατολισμένη στα έγγραφα – Wikipedia. Κάθε άρθρο αποτελείται από δεδομένα και ζεύγη κλειδιών-τιμών και συνδέσμους προς άλλα άρθρα, σχηματίζοντας έτσι ένα σύνολο σημασιολογικά δομημένων γραφημάτων.

Σε μία σύνθετη και πιο βαθιά ανάλυση της εικόνας 3.15, αξίζει να εστιάσουμε στο τίτλο του όρου (*label*), καθώς και στο μοναδικό αναγνωριστικό (*item identifier*). Φυσικά σημαντικό ενδιαφέρον παρουσιάζουν και οι υπόλοιποι όροι, όπως *property*, *rank*, *value* κ.α., ξεφεύγοντας από τον καίριο σκοπό μας, της απλής διασύνδεσης.



Εικόνα 3.15 – Σημαντικοί όροι στην Wikidata²³

Η διασύνδεση των δεδομένων των πρακτικών του Ελληνικού Κοινοβουλίου με τα αντίστοιχα του Wikidata γίνεται στο αρχικό στάδιο της δόμησής τους. Συγκεκριμένα κατά την διαδικασία της δημιουργίας δομημένων μορφών αρχείων xml, κάθε φορά που συναντάμε ένα νέο όνομα ομιλητή, περιηγούμαστε με αυτοματοποιημένο τρόπο στο διαδίκτυο και αναζητάμε στον ιστότοπο της Google το όνομα του ομιλητή, με την προσθήκη του όρου “wikidata”. Έπειτα αξιολογούμε κατάλληλα αν από το πρώτο αποτέλεσμα της αναζήτησης ανταποκρίνεται στην αναμενόμενη μορφή διεύθυνσης URL του Wikidata: <https://www.wikidata.org/wiki/Q>, δηλαδή ελέγχουμε αν υπάρχει αυτολεξεί ή παρόμοιος ο όρος που αναζητήσαμε στον ιστό του Wikidata. Οι επαληθευμένες διευθύνσεις URL που προκύπτουν χρησιμοποιούνται απευθείας στα δομημένα LegalDocML αρχεία, αλλά και αποθηκεύονται για μελλοντική χρήση.

Η διαδικασία που ακολουθούμε για την αποθήκευση και την εύκολη και γρήγορη αναζήτησή του σε μετέπειτα στάδιο, βασίζεται στην κατασκευή ενός νέου JSON αρχείου, στο

²³ Πηγή: <https://en.wikipedia.org/wiki/Wikidata#Concept>

όποιο περιέχονται ζεύγη ονομάτων και διευθύνσεων URL. Στην περίπτωση που το όνομα που αναζητήσαμε δεν επιστρέφει αποτέλεσμα στο Wikidata το όνομα τοποθετείται στο JSON αρχείο με την προσθήκη μιας απλής παύλας (“-”).

Διατρέχοντας όλα τα ονόματα ομιλητών από όλα τα κοινοβουλευτικά αρχεία, εξασφαλίζουμε ότι έχουμε πρόσβαση στους συνδέσμους Wikidata για τις οντότητες στο σύνολο δεδομένων μας. Αυτή η σύνδεση μας επιτρέπει να εμπλουτίσουμε τα δομημένα αρχεία μας με πρόσθετες πληροφορίες από το Wikidata, διευκολύνοντας την πιο ολοκληρωμένη και πλούσια σε περιεχόμενο αναπαράσταση γνώσης.

3.3.3 Μετατροπή Αρχείων XML σε RDF αρχεία

Σε αυτό το βήμα της εργασίας, αναλύουμε την μεθοδολογία για την σημασιολογική αναπαράσταση, με την μετατροπή των δομημένων αρχείων σε τριπλέτες μορφής RDF. Για τον σκοπό αυτό, χρησιμοποιούμε επικουρικά τις δυνατότητες της XSLT (eXtensible Stylesheet Language Transformations) σε συνδυασμό με την *rdflib*, μια δημοφιλή βιβλιοθήκη Python για την ενασχόληση με δεδομένα-αρχεία RDF.

Η είσοδος σε αυτό το υποσύστημα είναι τόσο τα δομημένα αρχεία, μορφής LegalDocML XML, που παράξαμε στο προηγούμενο στάδιο, όσο και αρχεία csv από το project που ασχολείται με τα ελληνικά κοινοβουλευτικά δεδομένα [9]. Επιλέξαμε αρχεία csv που περιέχουν πληροφορίες για πολιτικές θητείες και περιόδους κυβερνήσεων, γεγονός που αποσκοπεί στο να αυξήσει περαιτέρω το εύρος και το βάθος του συνόλου δεδομένων.

Πιο αναλυτικά, τα δομημένα αρχεία XML αποτελούνται από στοιχεία με ετικέτες που αντικατοπτρίζουν διάφορα μέρη του κειμένου, όπως τα απαραίτητα μεταδεδομένα, τα ονόματα των ομιλητών και το περιεχόμενο των ομιλιών τους. Αυτά απαιτούνται για την δημιουργία τριπλετών RDF, οι οποίες αναπαριστούν τις σχέσεις και τα μεταδεδομένα κάθε ομιλίας και των σχετικών ομιλητών.

Από την άλλη, τα csv αρχεία που έχουν πληροφορίες για την πολιτική θητεία και τους κυβερνητικούς ρόλους των βουλευτών τα μετατρέπουμε σε τριπλέτες RDF, αποτυπώνοντας τις πληροφορίες που σχετίζονται με τις πολιτικές θέσεις, τους ρόλους και τις περιόδους θητείας των ατόμων στο κοινοβούλιο. Τα δεδομένα αυτά είναι χρήσιμα για την κατανόηση του πολιτικού τοπίου και της κατανομής των αρμοδιοτήτων μεταξύ των μελών του κοινοβουλίου.

Στα πλαίσια αυτής εργασίας επιλέξαμε να μετατραπούν τα αρχεία rdf σε μορφή rdf/xml, όπως φαίνεται και στις εικόνες 3.16 όπου δίνονται ελάχιστες από τις τριπλέτες, μιας και υπολογίζεται πως σε ένα πρώτο στάδιο ο συνολικός αριθμός τριπλετών κυμαίνεται σε μεγάλο επταψήφιο (9εκ+). Συγκεκριμένα, παρακάτω φαίνονται μεμονωμένες rdf/xml περιπτώσεις, όπου πρόκειται ξεχωριστά στο (α) για τον 12° λόγο της συζήτησης την ημερομηνία 08/06/2018 (*debate_2018-06-8_1_speech_12*), στο (β) για τη βουλευτή Άννα-Μισέλ Ασημακοπούλου (*GRmember_996*), στο (γ) για πολιτικές θητείες και αξιώματα για την ίδια ομιλήτρια και τέλος στο (δ) για την πολιτική θητεία της ίδιας στον νομό Ιωαννίνων τα έτη 2012-2014 (*political_tenure_701*).

```

<rdf:Description rdf:about="https://purl.org/greekparldebates/debate_2018-06-08_1_speech_12">
  <dcterms:date rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2018-06-08</dcterms:date>
  <dcterms:language>gr</dcterms:language>
  <dcterms:isPartOf rdf:resource="https://purl.org/greekparldebates/akn/gr/debate/2018-06-08"/>
  <greek_lp:hasSubsequent rdf:resource="https://purl.org/greekparldebates/debate_2018-06-08_1_speech_13"/>
  <greek_lp:speaker rdf:resource="https://purl.org/greekparldebates/GRmember_996"/>
  <greek_lp:spokenText xml:lang="el">Αυτά έχουν καταβληθεί. Πολύ ωραία. Δέχομαι την απάντηση, αλλά θα ήθελα
  να ξέρω το εξής: Υπάρχουν ενέργειες για το μέλλον; Υπάρχει κάτι το οποίο θα μας εξασφαλίσει ότι ο κόσμος
  δεν θα ξαναβρίσκεται εκεί; Ό,τι και να συζητάμε, όσο και να συγκρούμαστε εδώ μέσα, αυτός που κάθεται στην
  ουρά θέλει να έχει μια απάντηση για το τι θα γίνει όταν θα πάει να ξανσπληρώσει τον λογαριασμό του στον
  ΕΑΤΑ.
  Ευχαριστώ.</greek_lp:spokenText>
</rdf:Description>

```

Εικόνα 3.16 (α)

```

<rdf:Description rdf:about="https://purl.org/greekparldebates/GRmember_996">
  <foaf:name>anna-misel_asimakopoulou</foaf:name>
  <foaf:gender>female</foaf:gender>
  <owl:sameAs rdf:resource="https://www.wikidata.org/wiki/Q16329215"/>
</rdf:Description>

```

Εικόνα 3.16 (β)

```

<rdf:Description rdf:about="https://purl.org/greekparldebates/GRmember_996">
  <greek_lp:PoliticalTenure rdf:resource="https://purl.org/greekparldebates/political_tenure_701"/>
  <greek_lp:PoliticalTenure rdf:resource="https://purl.org/greekparldebates/political_tenure_702"/>
  <greek_lp:PoliticalTenure rdf:resource="https://purl.org/greekparldebates/political_tenure_703"/>
</rdf:Description>

```

Εικόνα 3.16 (γ)

```

<rdf:Description rdf:about="https://purl.org/greekparldebates/political_tenure_701">
  <greek_lp:beginning rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2012-06-17</greek_lp:beginning>
  <greek_lp:end rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2014-12-31</greek_lp:end>
  <greek_lp:Party rdf:resource="https://purl.org/greekparldebates/nea_dimokratia"/>
  <greek_lp:administrative_region>ioanninon</greek_lp:administrative_region>
  <rdfs:label xml:lang="en">Political Tenure 701</rdfs:label>
</rdf:Description>

```

Εικόνα 3.16 (δ)

Εικόνα 3.16 – Τριπλέτες RDF όπως αποτυπώνονται στα rdf/xml αρχεία

Τέλος, εμβαθύνοντας στην αναπαράσταση του σημασιολογικού μοντέλου (εικόνα 3.17) βλέπουμε ένα μέρος του συνόλου δεδομένων των πρακτικών του Ελληνικού Κοινοβουλίου. Με την πρώτη ματιά, αυτή η εικόνα μπορεί να φαίνεται σαν ένα πολύπλοκο δίκτυο κόμβων και συνδέσεων, αλλά στην ουσία συμπυκνώνει τους λόγους, τους ομιλητές και τις πολιτικές θητείες που αναφέρονται στην εικόνα 3.9.

3.3.3.1 RDF διασύνδεση με Wikidata

Για τη διασύνδεση των δεδομένων των αρχείων rdf με την Wikidata αξιοποιούμε κατάλληλα το json αρχείο που δημιουργήσαμε για τους ομιλητές στο στάδιο κατασκευής και διασύνδεσης των πρώτων δομημένων αρχείων, μορφής LegalDocML. Με αυτόν τον τρόπο επιτυγχάνουμε διασύνδεση των πληροφοριών που αφορά τους ομιλητές με γρήγορο και απλό τρόπο, καθώς δεν απαιτείται περιττή αναζήτηση στο διαδίκτυο, αλλά μια αξιοποίηση των ήδη υπάρχουσών πληροφοριών από το json αρχείο.

Επιπλέον, προκειμένου να επιτευχθεί όσο τον δυνατόν υψηλότερη διασύνδεση με την Wikidata επιλέξαμε να επαναλάβουμε την διαδικασία αναζήτησης μέσω Google και για τα κόμματα και για τα υπουργικά αξιώματα. Έτσι δημιουργήσαμε τριπλέτες rdf για τα κόμματα συνδέοντας κάθε κόμμα με το αντίστοιχο url προς την βάση Wikidata. Όσον αφορά τις υπουργικές θέσεις επιλέξαμε να συνδεθεί η κάθε θέση με το αντίστοιχο υπουργείο γενικότερα και όχι με την αντίστοιχη συγκεκριμένη θέση.

Ενδεικτικά ένα στοχευμένο παράδειγμα που αποτυπώνει την διασύνδεση βρίσκεται στην Εικόνα 3.16 (β), όπου εμφανίζεται η οντότητα που αντιπροσωπεύει τον ομιλητή με *id* = *GRmember_996*, δηλαδή τη διεύθυνση 'https://purl.org/greekparldebates/GRmember_996'. Πέρα από τις γνωστές ιδιότητες (όνομα και φύλο), η '<owl:sameAs>' χρησιμοποιείται για τη δημιουργία σύνδεσης μεταξύ αυτής της συγκεκριμένης οντότητας και μιας αντίστοιχης καταχώρησης στο Wikidata, με URL "<https://www.wikidata.org/wiki/Q16329215>"

3.3.3.2 RDF με κυβερνητικά δεδομένα από επίσημες πηγές

Ένα ακόμα βασικό στοιχείο της βελτίωσης της αναπαράστασης της γνώσης είναι η ενσωμάτωση κυβερνητικών δεδομένων από επίσημες πηγές, συμπεριλαμβανομένων λεπτομερειών σχετικά με τις κυβερνήσεις και τους υφισταμένους τους. Ο επίσημος δικτυακός τόπος της Γενικής Γραμματείας Νομικών και Κοινοβουλευτικών Υποθέσεων²⁴, μια συλλογή αξιόπιστων νομικών και κοινοβουλευτικών πληροφοριών, είναι μια τέτοια ανεκτίμητη πηγή. Αποστολή της Γενικής Γραμματείας Νομικών και Κοινοβουλευτικών Θεμάτων είναι η διασφάλιση της συνοχής και του συντονισμού της διαδικασίας παραγωγής νόμου, η αποτελεσματική εφαρμογή των αρχών και εργαλείων της Καλής Νομοθέτησης, καθώς και η υποστήριξη του Υπουργικού Συμβουλίου και των συλλογικών κυβερνητικών οργάνων.

Ο ιστότοπος της Γενικής Γραμματείας λειτουργεί ως κομβικό σημείο για κοινοβουλευτικά και νομικά θέματα, καθώς περιέχει πληθώρα πληροφοριών σχετικά με κυβερνητικά σχήματα, νομικές εξελίξεις και διοικητικές διαδικασίες. Χρησιμοποιώντας αυτή την επίσημη πηγή, μπορούμε να αποκτήσουμε αξιόπιστες πληροφορίες σχετικά με τις κυβερνήσεις, την ιεραρχία, τα καθήκοντα και τις λειτουργίες τους. Ως αποτέλεσμα, ενισχύουμε τις τριπλέτες RDF με στοιχεία και λεπτομερείς πληροφορίες σχετικά με τα κυβερνητικά πλαίσια, τις νομοθετικές διαδικασίες και τους βουλευτές. Με άλλα λόγια, εκμεταλλευόμαστε δεδομένα που αφορούν κυβερνήσεις συμπεριλαμβανομένων, όχι μόνο των ονομάτων των πρωθυπουργών, αλλά ακόμα και ολόκληρης της υπουργικής ηγεσίας. Έτσι είναι αναμενόμενο να παρουσιάζεται μια πιο ολοκληρωμένη εικόνα των πολύπλοκων αλληλεπιδράσεων μεταξύ των κοινοβουλευτικών διαδικασιών, των κυβερνητικών οργάνων και των ανθρώπων που επηρεάζουν τις νομοθετικές αποφάσεις.

²⁴ https://gslegal.gov.gr/?page_id=776&sort=time

Επίσης, χρήσιμες πληροφορίες εξάγουμε και από τον επίσημο ιστότοπο του Ελληνικού Κοινοβουλίου²⁵ ενισχύοντας την αναπαράσταση της γνώσης μας να περιλαμβάνει περεταίρω πληροφορίες για τους βουλευτές. Πιο συγκεκριμένα, αντλούμε δεδομένα για την κοινοβουλευτική θητεία του εκάστοτε, παραθέτοντας στοιχεία για την περίοδο, την περιφέρεια, αλλά και την κοινοβουλευτική ομάδα στην οποία ανήκει. Με αυτό το χαρακτηριστικό, είμαστε πλέον σε θέση να καταγράψουμε πλήρως δεδομένα για τους βουλευτές και την κοινοβουλευτική ιστορία τους.

3.3.4 Μετατροπή Αρχείων XML LegalDocML σε TEI ParlaMint

Στην τελευταία αυτή παράγραφο, παρουσιάζουμε τον τρόπο που ακολουθήσαμε για την δημιουργία νέων δομημένων κοινοβουλευτικών αρχείων της μορφή (TEI) ParlaMint.

Για τον σκοπό αυτό, χρησιμοποιήσαμε ως αρχικά δεδομένα τα ήδη δομημένα αρχεία, μορφής LegalDocML, τα οποία αποτελούν ακριβές και δομημένη αναπαράσταση των πρακτικών της Βουλής.

Η διαδικασία μετατροπής που αναπτύξαμε είναι ιδιαίτερα απλή και περιλαμβάνει την αντιστοίχιση των στοιχείων και των χαρακτηριστικών από τα LegalDocML αρχεία στα αντίστοιχα στοιχεία του TEI ParlaMint, διασφαλίζοντας τη διατήρηση της ακεραιότητας και του νοήματος των δεδομένων. Ως άξονας σε αυτή τη διαδικασία μετατροπής χρησιμεύει η γλώσσα XSLT. Στην παρούσα εφαρμογή, το αρχείο που περιλαμβάνει τους κανόνες για την αντιστοίχιση αντλήθηκε από το δημόσιο repository της δημιουργού ομάδας Clarin, το `akn2tei.xml`²⁶. Βασιζόμενοι σε αυτό το αρχείο και με την χρήση της γλώσσας XSLT, αντιστοιχίσαμε με ακρίβεια τα στοιχεία και τις ετικέτες των αρχείων και τα μετατρέψαμε στα τελικά δομημένα αρχεία της μορφής ParlaMint. Έτσι, τα μοτίβα στο LegalDocML XML εντοπίζονται και μεταφράζονται στη δομημένη μορφή του TEI ParlaMint.

Ενδεικτικά, ένα χαρακτηριστικό παράδειγμα της σύνταξης ενός δομημένου αρχείου, που βασίζεται στην μορφή του ParlaMint βρίσκεται στην εικόνα 3.18.

```
<u xml:id="debate_2018-06-08_1_speech_12" who="anna-misel_asismakopoulou">
  <note type="speaker">ΑΝΝΑ-ΜΙΣΕΛ ΑΣΗΜΑΚΟΠΟΥΛΟΥ</note>
  <seg>Αυτά έχουν καταβληθεί. Πολύ ωραία. Δέχομαι την απάντηση, αλλά θα ήθελα να
  ξέρω το εξής: Υπάρχουν ενέργειες για το μέλλον; Υπάρχει κάτι το οποίο θα μας
  εξασφαλίσει ότι ο κόσμος δεν θα ξαναβρίσκεται εκεί; Ό,τι και να συζητάμε, όσο
  και να συγκρούμαστε εδώ μέσα, αυτός που κάθεται στην ουρά θέλει να έχει μια
  απάντηση για το τι θα γίνει όταν θα πάει να ξαναπληρώσει τον λογαριασμό του
  στον ΕΛΤΑ.</seg>
  <seg>Ευχαριστώ.</seg>
</u>
<u xml:id="debate_2018-06-08_1_speech_13" who="dimitrios_kremastinos">
  <note type="speaker">Δημήτριος Κρεμαστινός</note>
  <seg>Ευχαριστούμε.</seg>
  <seg>Κύριε Υπουργέ, έχετε τον λόγο.</seg>
</u>
<u xml:id="debate_2018-06-08_1_speech_14" who="nikolaos_pappas">
  <note type="speaker">ΝΙΚΟΛΑΟΣ ΠΑΠΠΑΣ</note>
  <seg>Ευχαριστώ, κύριε Πρόεδρε.</seg>
  <seg>Κυρία Ασημακοπούλου, νομίζω ότι δεν με παρακολουθήσατε.</seg>
</u>
```

Εικόνα 3.18 – Απόσπασμα από κύριο μέρος ενός αρχείου ParlaMint

Όσον αφορά την τελευταία εικόνα (3.18), είναι χαρακτηριστική η ιδιόμορφη σύνταξη του ParlaMint αποσπάσματος. Κάθε λόγος περικλείεται εξωτερικά από την ετικέτα `<u>`, με χαρακτηριστικά να έχει το `xml:id`, αναγνωριστικό id για κάθε λόγο και το `who`, αναγνωριστικό του εκάστοτε ομιλητή. Στην συνέχεια υπάρχουν δύο ακόμα ετικέτες, `<note>`

²⁵ <https://www.hellenicparliament.gr/Vouleftes/Diatelesantes-Vouleftes-Apo-Ti-Metapolitefsi-Os-Simera/>

²⁶ <https://github.com/clarin-eric/parla-clarin/blob/master/Examples/AkomaNtoso/akn2tei.xml>

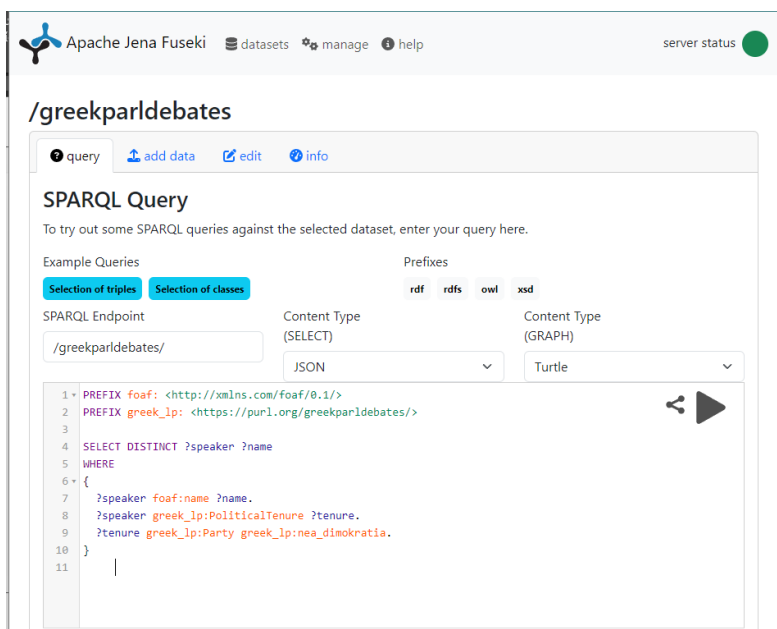
και <seg>. Στο περιεχόμενο της ετικέτας <note> βρίσκεται αυτούσιο το όνομα του ομιλητή, και μέσα στην ετικέτα <seg> περικλείονται τα τμήματα του λόγου.

3.4 Πρόσβαση και Αναζήτηση δεδομένων RDF - Apache Jena Fuseki

Σε επόμενο βήμα βρίσκεται η παροχή συστήματος για την πρόσβαση στην δομημένη σημασιολογική αναπαράσταση των κοινοβουλευτικών συζητήσεων. Για αυτό το σκοπό αξιοποιούμε τα αρχεία RDF που είναι εμπλουτισμένα με δεδομένα για κοινοβουλευτικές συνεδριάσεις, ομιλίες, ομιλητές, πολιτικές θητείες και κυβερνητικούς ρόλους.

Επιλέξαμε να χρησιμοποιήσουμε το σύστημα Apache Jena Fuseki, ένα σύστημα βάσεων δεδομένων RDF, για να επιτρέψουμε την απλή πρόσβαση και την υποβολή ερωτημάτων στα δεδομένα RDF. Δημιουργώντας την βάση δεδομένων, αναρτούμε όλα τα rdf αρχεία που έχουμε κατασκευάσει και είναι απαραίτητα για την εργασία. Αφού ολοκληρωθεί η διαδικασία αυτή, μπορούμε να διατυπώσουμε ακριβή ερωτήματα και να λάβουμε συγκεκριμένες πληροφορίες από τα τελικά συνδεδεμένα δεδομένα, χρησιμοποιώντας την SPARQL, τη γλώσσα ερωτημάτων για RDF.

Έχει δημιουργηθεί ένας server, όπου μέσω της ιστοσελίδας <http://debates.dslab.ece.ntua.gr:3032/>, μπορεί ο καθένας να έχει πρόσβαση στα τελικά αρχεία μας και να υποβάλει ερωτήματα SPARQL, επάνω στις ήδη υπάρχουσες τριπλέτες. Το περιβάλλον χρήσης φαίνεται παρακάτω όπου στην πρώτη εικόνα φαίνεται η θέση που θέτουμε ερωτήματα (Εικόνα 3.19), και στην δεύτερη εικόνα είναι ορατό το αποτέλεσμα ενός ερωτήματος σε μορφή πίνακα (Εικόνα 3.20).



Εικόνα 3.19 – Περιβάλλον υποβολής ερωτημάτων SPARQL

Table

Response

534 results in 0.151 seconds

Simple view

Ellipse

Filter query results

Page size: 50

speaker	name
1 <https://purl.org/greekparldebates/GRmember_392>	parthena_fountoukidou-theodoridou
2 <https://purl.org/greekparldebates/GRmember_31>	xristos_koskinas
3 <https://purl.org/greekparldebates/GRmember_743>	georgios_papageorgiou
4 <https://purl.org/greekparldebates/GRmember_2541>	mixail_xalkidis
5 <https://purl.org/greekparldebates/GRmember_4373>	sumeon_kedikoglou
6 <https://purl.org/greekparldebates/GRmember_1193>	athanasios_zempiis
7 <https://purl.org/greekparldebates/GRmember_1594>	georgios_paulakakis
8 <https://purl.org/greekparldebates/GRmember_2543>	grigorios_apostolakis
9 <https://purl.org/greekparldebates/GRmember_1105>	georgios_kasapidis
10 <https://purl.org/greekparldebates/GRmember_1240>	konstantinos_kollias
11 <https://purl.org/greekparldebates/GRmember_217>	konstantinos_georgolios
12 <https://purl.org/greekparldebates/GRmember_958>	olga_kefalogianni
13 <https://purl.org/greekparldebates/GRmember_354>	ioannis_lampropoulos
14 <https://purl.org/greekparldebates/GRmember_1201>	konstantinos_skrekas
15 <https://purl.org/greekparldebates/GRmember_279>	athanasios_nakos
16 <https://purl.org/greekparldebates/GRmember_118>	anna_mpenaki-psarouda
17 <https://purl.org/greekparldebates/GRmember_403>	theofilos_leontaridis
18 <https://purl.org/greekparldebates/GRmember_4576>	georgios_stamatis
19 <https://purl.org/greekparldebates/GRmember_2412>	nikolaos_arguopoulos
20 <https://purl.org/greekparldebates/GRmember_2239>	ioannis_panagiotidis
21 <https://purl.org/greekparldebates/GRmember_46>	dimitrios_kostopoulos
22 <https://purl.org/greekparldebates/GRmember_412>	theodoros_damianos

Εικόνα 3.20 – Δείγμα απάντησης ερωτημάτων SPARQL σε μορφή πίνακα

3.5 Σχολιασμός Διαδικασίας Μετατροπής

Γενικά, η ομαλή μετατροπή εγγράφων από μια μορφή σε μια άλλη είναι απαραίτητη για την διαχείριση μεγάλου όγκου αρχείων. Ωστόσο, υπάρχουν αρκετές δυσκολίες, που εμποδίζουν μια αυτοματοποιημένη διαδικασία να εκτελεί αυτή τη μετατροπή, κάποιες από τις οποίες συναντήθηκαν και εδώ.

Στην παρούσα εργασία, σε πρώτο στάδιο, καλούμαστε να κατασκευάσουμε ένα σύστημα στο οποίο εισάγονται αδόμητα αρχεία, μορφής doc(x)/txt, και μετατρέπονται σε δομημένα αρχεία, και μορφής LegalDocML. Η πολυπλοκότητα ενός τέτοιου συστήματος σε συνάρτηση με τον ανθρώπινο παράγοντα συγγραφής των πρακτικών αποτέλεσαν προκλήσεις στην υλοποίηση του συστήματος μας.

Στη συγκεκριμένη εργασία, καταφέραμε ένα ποσοστό επιτυχίας μετατροπής αρχείων που αγγίζει το 96%, καθώς μετατρέψαμε τα 5.571 από τα 5.801 αρχικά αρχεία της βάσης δεδομένων μας. Ένα βασικό εμπόδιο το οποίο δεν επέτρεψε την αποφυγή σφαλμάτων και την 100% επιτυχία αυτού του συστήματος είναι τα συντακτικά και τυπογραφικά λάθη. Πιο συγκεκριμένα, το σύστημα διαβάζει το κείμενο μέσω ενός τυποποιημένου μοντέλου, το οποίο αναγνωρίζει και αντιστοιχίζει τις πληροφορίες των αρχικών κειμένων που εισάγονται. Από το σύστημα αυτό ωστόσο προέκυψαν κάποια «ελλατωματικά» αρχεία, στα οποία παρατηρούμε πως ο συντάκτης έχει ξεφύγει από τους βασικούς συντακτικούς κανόνες του μοντέλου, είτε παραλείποντας κάποια βασικά εισαγωγικά στοιχεία είτε παραθέτοντας κάποια νέα και ιδιόμορφα.

Τα τυπογραφικά λάθη, παρά το γεγονός ότι φαίνονται ασήμαντα, μπορούν να παρεμποδίσουν σοβαρά την διαδικασία δόμησης και διασύνδεσης των κοινοβουλευτικών συζητήσεων. Η παρερμηνεία μπορεί να προκύψει από ανορθόγραφες λέξεις, ακατάλληλη στίξη ή ακανόνιστη μορφοποίηση.

Επομένως εξαιτίας του μικρού ποσοστού σφάλματος (~4%), η περαιτέρω ενασχόληση με τα συγκεκριμένα σφάλματα αποτέλεσε δευτερεύοντα στόχο διότι η επίλυσή τους θα ήταν ιδιαίτερα χρονοβόρα και εξαντλητική διαδικασία και θα έπρεπε να εξετάσουμε κάθε περίπτωση ξεχωριστά.

Τέλος, αξίζει να σχολιάσουμε πως η εξαγωγή δομημένων αρχείων μορφής ParlaMint πραγματοποιήθηκε χωρίς κανένα εμπόδιο. Πιο συγκεκριμένα, η τυποποιημένη δομή των αρχείων LegalDocML ήταν ένα από τα σημαντικά στοιχεία για την επιτυχία αυτής της μετατροπής, μιας και αυτά ακολουθούσαν μια ενιαία μορφή, η οποία έκανε την εξαγωγή και τον χειρισμό των δεδομένων πολύ απλούστερη. Έτσι, η σαφώς καθορισμένη δομή αυτών των αρχείων κατέστησε απλή την ανάκτηση τιμών για τη μετατροπή σε ParlaMint. Τα βασικά δεδομένα μπορούσαν να εξαχθούν προγραμματιστικά χάρη στη χρήση τυποποιημένων ετικετών και χαρακτηριστικών, γεγονός που εξάλειψε την πιθανότητα σφαλμάτων που μερικές φορές προκύπτουν από συγκεχυμένες ή ασυνεπείς μορφές δεδομένων.

4

Μελέτη Περιπτώσεων και Στατιστικά Δεδομένα


Σε αυτήν την ενότητα, παρουσιάζουμε περιπτώσεις μελέτης στα δομημένα αρχεία, και στις πληροφορίες των σημασιολογικών κοινοβουλευτικών αρχείων. Από τα πρώτα εξάγουμε θεματικούς τίτλους και από τα δεύτερα αντλούμε συγκεκριμένα δεδομένα μέσα από σημασιολογικά ερωτήματα.

Τέλος, παρουσιάζουμε κάποια συνολικά στατιστικά που προέκυψαν από τα δομημένα αρχεία των κοινοβουλευτικών συζητήσεων.

4.1 Περίπτωση μελέτης RDF

Σε αυτήν την παράγραφο, παρουσιάζουμε παραδείγματα περιπτώσεων στα τελικά σημασιολογικά αρχεία των κοινοβουλευτικών συζητήσεων. Πιο συγκεκριμένα, χρησιμοποιώντας ερωτήματα SPARQL εντός του συστήματος διαχείρισης δεδομένων Apache Jena Fuseki, πλοηγούμαστε και αλληλεπιδρούμε με τα δομημένα δεδομένα των τριπλετών RDF, εξασφαλίζοντας πρόσβαση σε σχετικές λεπτομέρειες από τις κοινοβουλευτικές συζητήσεις. Αυτή η ολοκληρωμένη προσέγγιση μας επιτρέπει να αξιοποιήσουμε αποτελεσματικά τη δύναμη των σημασιολογικών τεχνολογιών και να κάνουμε ερωτήματα στα δεδομένα του Ελληνικού Κοινοβουλίου που είναι αποθηκευμένα στο σύστημα μας.

Έστω, ότι επιθυμούμε να εντοπίσουμε ποιες ημερομηνίες μίλησε ένας πολιτικός μέσα στο Ελληνικό Κοινοβούλιο. Κάνοντας την σύμβαση ότι το χρονικό πεδίο ενδιαφέροντος είναι όλο αυτό που μελετάμε (1989 – 2023) και επιλέγοντας ενδεικτικά τον βουλευτή “Αλέξη/Αλέξιο Τσίπρα”, διατυπώνουμε κατάλληλα το ερώτημα που θέλουμε στην γλώσσα ερωτημάτων SPARQL. Όπως φαίνεται και παρακάτω στο δομημένο και ολοκληρωμένο ερώτημα, έχει επιλεγθεί η αναζήτηση του ομιλητή (speaker) με όνομα “*alexios_tsipras*”/“*alexis_tsipras*”.


Apache Jena Fuseki
datasets
manage
help
server status

/greekparldebates

query
add data
edit
info

SPARQL Query

To try out some SPARQL queries against the selected dataset, enter your query here.

Example Queries

Selection of triples **Selection of classes**

Prefixes

rdf **rdfs** **owl** **xsd**

SPARQL Endpoint

/greekparldebates/

Content Type (SELECT)

JSON

Content Type (GRAPH)

Turtle

```

1 PREFIX greek_lp: <https://purl.org/greekparldebates/>
2 PREFIX dcterms: <http://purl.org/dc/terms/>
3 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
4
5 SELECT DISTINCT ?date
6 WHERE
7 {
8   ?speech greek_lp:speaker ?speaker.
9   ?speech dcterms:date ?date.
10  ?speaker foaf:name ?name.
11  FILTER (?name = "alexis_tsipras" || ?name = "alexios_tsipras").
12 }
13

```

Εικόνα 4.1 – Παράδειγμα χρήσης – ερώτημα 1 με «Αλέξιος/Αλέξης Τσίπρα»

Στο συγκεκριμένο ερώτημα, με αντικείμενο μελέτης τον Αλέξη Τσίπρα, το αποτέλεσμα είναι ένας πίνακας με όλες τις ημερομηνίες που έχει μιλήσει. Φυσικά, πέρα από το σύνολο των ημερομηνιών, εμφανίζεται και ο αριθμός με το πλήθος των ημερομηνιών. Να σημειωθεί το γεγονός πως εφόσον έχουμε χρησιμοποιήσει τον όρο “*DISTINCT*” στο ερώτημα, αναμένουμε το αποτέλεσμα να μην περιέχει διπλότυπες ημερομηνίες.

Table
Response
323 results in 0.218 seconds

Simple view
Ellipse

Page size: 50

	date
1	"2016-05-08" <http://www.w3.org/2001/XMLSchema#date>
2	"2019-07-22" <http://www.w3.org/2001/XMLSchema#date>
3	"2021-11-19" <http://www.w3.org/2001/XMLSchema#date>
4	"2015-10-16" <http://www.w3.org/2001/XMLSchema#date>
5	"2016-03-29" <http://www.w3.org/2001/XMLSchema#date>
6	"2021-11-12" <http://www.w3.org/2001/XMLSchema#date>
7	"2015-10-30" <http://www.w3.org/2001/XMLSchema#date>
8	"2019-07-20" <http://www.w3.org/2001/XMLSchema#date>
9	"2021-11-22" <http://www.w3.org/2001/XMLSchema#date>
10	"2015-11-17" <http://www.w3.org/2001/XMLSchema#date>
11	"2022-05-26" <http://www.w3.org/2001/XMLSchema#date>
12	"2021-12-18" <http://www.w3.org/2001/XMLSchema#date>
13	"2022-01-29" <http://www.w3.org/2001/XMLSchema#date>
14	"2022-01-30" <http://www.w3.org/2001/XMLSchema#date>
15	"2022-05-12" <http://www.w3.org/2001/XMLSchema#date>
16	"2021-12-01" <http://www.w3.org/2001/XMLSchema#date>
17	"2022-05-10" <http://www.w3.org/2001/XMLSchema#date>
18	"2022-03-11" <http://www.w3.org/2001/XMLSchema#date>
19	"2022-02-15" <http://www.w3.org/2001/XMLSchema#date>
20	"2022-03-23" <http://www.w3.org/2001/XMLSchema#date>

Εικόνα 4.2 – Παράδειγμα χρήσης – αποτέλεσμα στο ερώτημα 1 με «Αλέξιος/Αλέξης Τσίπρα»

Σε ένα δεύτερο παράδειγμα, θέτουμε θέμα μελέτης το πόσες φορές μίλησε ένας βουλευτής από ένα συγκεκριμένο κόμμα για μία συγκεκριμένη χρονική περίοδο. Έστω πως επιλέγουμε και ενδιαφερόμαστε για βουλευτές του κόμματος της «Νέας Δημοκρατίας» για το έτος 2022. Ομοίως με πριν διατυπώνουμε το SPARQL ερώτημα, όπου χαρακτηριστικά φαίνεται ο ενδιαφερόμενος όρος “*nea_dimokratia*”, αλλά και το χρονικό διάστημα.

The screenshot shows the Apache Jena Fuseki web interface. At the top, there's a navigation bar with 'datasets', 'manage', and 'help' links, and a 'server status' indicator. Below this, the URL '/greekparldebates' is shown. The main area is titled 'SPARQL Query' and contains a text area with the following query:

```

1 PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
2 PREFIX greek_lp: <https://purl.org/greekparldebates/>
3 PREFIX dct: <http://purl.org/dc/terms/>
4 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
5
6 SELECT DISTINCT ?speaker ?name ?date
7 WHERE
8 {
9   ?speech greek_lp:speaker ?speaker.
10  ?speech dct:date ?date.
11
12  ?speaker foaf:name ?name.
13  ?speaker greek_lp:PoliticalTenure ?tenure.
14  ?tenure greek_lp:Party greek_lp:nea_dimokratia
15  FILTER (?date >="2022-01-01"^^xsd:date && ?date <="2022-12-31"^^xsd:date).
16 }
17

```

The interface also shows options for 'Example Queries', 'Prefixes' (rdf, rdfs, owl, xsd), 'SPARQL Endpoint' (/greekparldebates/), 'Content Type (SELECT)' (JSON), and 'Content Type (GRAPH)' (Turtle).

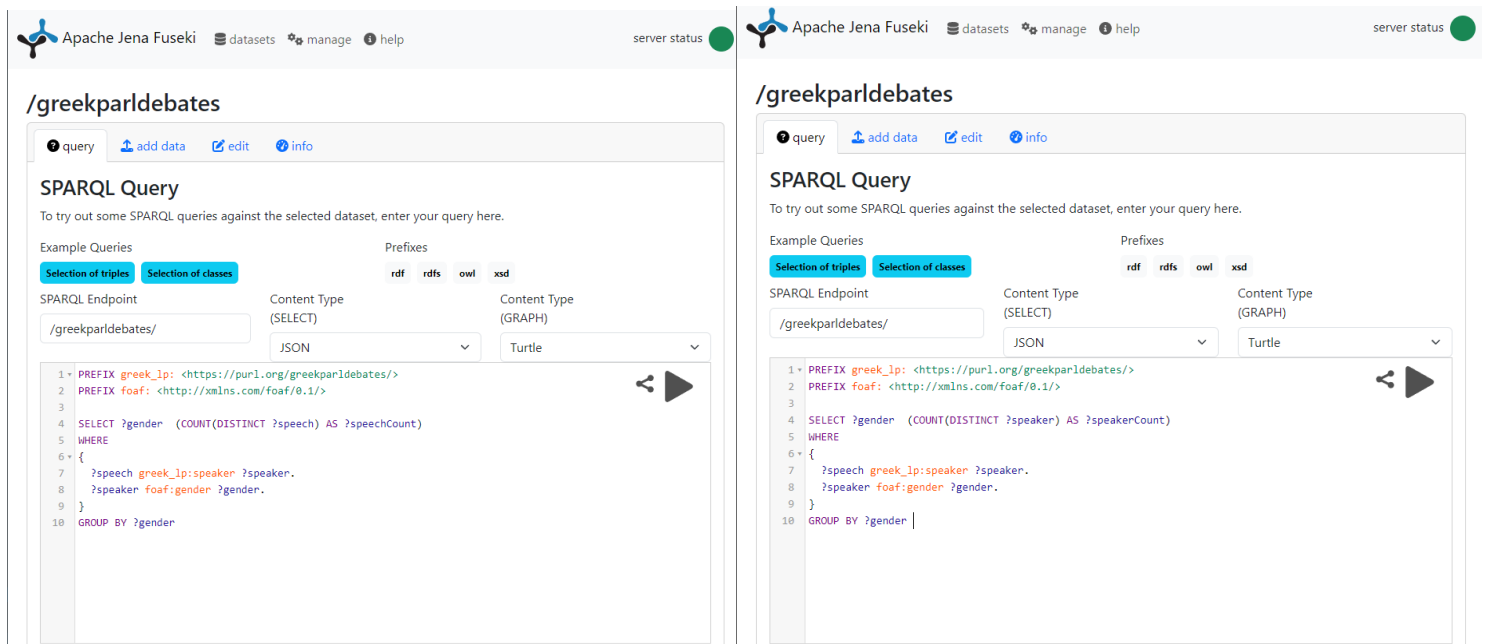
Εικόνα 4.3 – Παράδειγμα χρήσης – ερώτημα 2 με βουλευτές Νέας Δημοκρατίας

Το αποτέλεσμα του ερωτήματος φαίνεται στην εικόνα 4.4, όπου καταγράφονται 2.353 μοναδικές εγγραφές, όπου διακρίνεται το μοναδικό αναγνωριστικό κάθε ομιλητή (στήλη *speaker*), το όνομά του όπως είναι αποθηκευμένο στα σημασιολογικά αρχεία RDF (στήλη *name*) και τέλος η κάθε ημερομηνία (στήλη *date*).

Table			Response	2353 results in 15.249 seconds	Simple view	Ellipse	Filter query results	Page size: 50		
speaker	name	date								
1	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
2	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
3	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
4	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
5	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
6	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
7	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
8	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
9	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
10	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
11	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
12	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
13	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								
14	<https://purl.org/greekparldebates/GRmember_520>	maurodis_boridis								

Εικόνα 4.4 – Παράδειγμα χρήσης – αποτέλεσμα στο ερώτημα 2 με βουλευτές της Νέας Δημοκρατίας

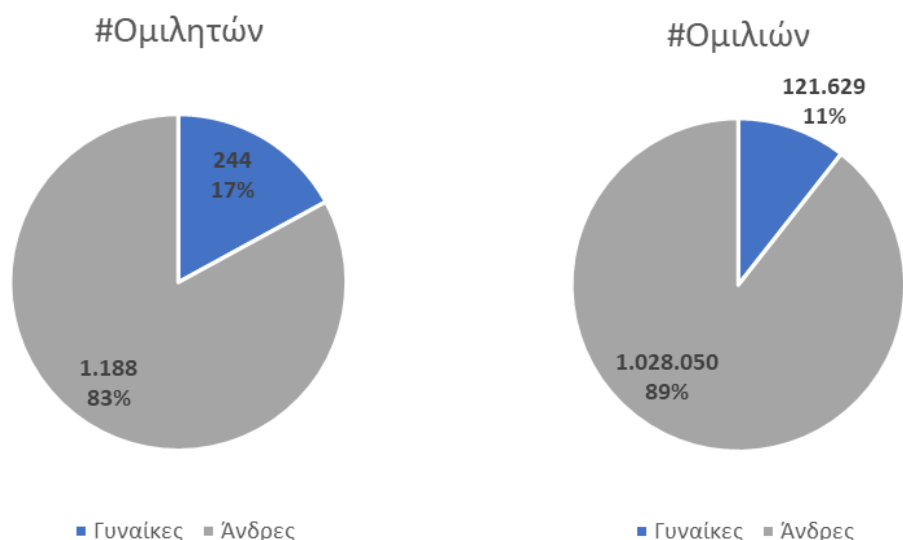
Ένα άλλο ένα δεδομένο που εξάγουμε, αφορά την εκπροσώπηση των δύο φύλων στο Ελληνικό Κοινοβούλιο. Πιο συγκεκριμένα υπάρχει μια σαφή διαφορά μεταξύ του τρόπου με τον οποίο εκπροσωπούνται οι άνδρες και οι γυναίκες μεταξύ των ομιλητών σε ένα δεδομένο περιβάλλον. Με μόλις 244 ομιλήτριες σε ένα σύνολο 1.432 ομιλητών, ή περίπου 17% του συνόλου, υπάρχουν σημαντικά λιγότερες γυναίκες ομιλητές εντός του κοινοβουλευτικού χώρου. Συγκριτικά, οι άνδρες ομιλητές αντιπροσωπεύουν 1.188 του συνόλου, ποσοστό αρκετά μεγαλύτερο. Κατ' επέκταση, το χάσμα μεταξύ των δύο φύλων εξακολουθεί να υπάρχει στο πλήθος των λόγων τους, όπου το θηλυκό φύλο έχει εκφέρει 121.629 λόγους, σε σχέση με το αρσενικό που έχει εξαψήφιο, δηλαδή 1.028.050. Όλα αυτά έχουν συγκεντρωθεί στον Πίνακα 4.1, ο οποίος προέκυψε από το συνδυασμό των δύο ερωτημάτων της εικόνας 4.5. Στο αριστερό μέρος της εικόνας, καταγράφουμε το ερώτημα που επιστρέφει το πλήθος των ομιλιών χωρισμένο ανά φύλο. Ενώ το δεξί, αφορά τον αριθμό των ομιλητών ανά φύλο. Η αποτύπωση των στατιστικών δεδομένων ολοκληρώνεται με τα διαγράμματα της εικόνας 4.6.



Εικόνα 4.5 – Ερωτήματα ομιλιών και ομιλητών ανά φύλο

ΠΙΝΑΚΑΣ 4.1 – Αριθμός ομιλητών και ομιλιών ανά φύλο στις συζητήσεις του Ελληνικού Κοινοβουλίου

	Γυναίκες	Άνδρες
#Ομιλιών	121.629	1.028.050
#Ομιλητών	244	1.188



Εικόνα 4.6 – Διαγράμματα ομιλητών και ομιλιών ανά φύλο στις συζητήσεις του Ελληνικού Κοινοβουλίου

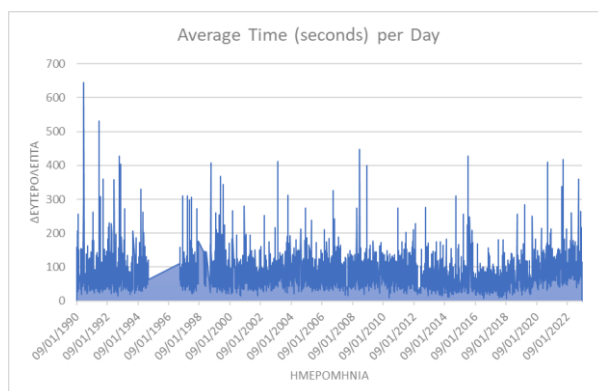
Σε συνέχεια του προηγούμενου, ενδιαφέρον παρουσιάζει η αναλογία των δύο φύλων ανά πολιτική δύναμη. Εξετάζοντας πιο προσεκτικά τον πίνακα 4.2, γίνεται σαφές ότι το χάσμα μεταξύ των δύο φύλων επηρεάζει πρακτικά όλα τα πολιτικά κόμματα. Οι διαφορές είναι εντυπωσιακές, δείχνοντας ότι η υποεκπροσώπηση των γυναικών είναι ένα ευρέως διαδεδομένο ζήτημα και δεν επηρεάζει μόνο μερικές εξειδικευμένες πολιτικές ομάδες. Ο μέσος αριθμός των γυναικών ομιλητών στην πλειοψηφία των κόμματος δεν υπερβαίνει το 25%, γεγονός που είναι ιδιαίτερα σημαντικό.

ΠΙΝΑΚΑΣ 4.2 – Αριθμός ομιλητών-βουλευτών ανά κόμμα και φύλο στις συζητήσεις του Ελληνικού Κοινοβουλίου

Πολιτική Δύναμη	#Γυναίκες	#Άνδρες
Ανεξάρτητοι (εκτός κόμματος)	27	118
Ανεξάρτητοι Δημοκρατικοί Βουλευτές	4	14
Ανεξάρτητοι Έλληνες - Πάνος Καμμένος	7	19
Ανεξάρτητοι Έλληνες- Εθνική Πατριωτική Δημοκρατική Συμμαχία	3	14
Δημοκρατική Ανανέωση	0	2
Δημοκρατική Αριστερά	5	13
Δημοκρατική Συμπράταξη (Πανελλήνιο Σοσιαλιστικό Κίνημα - Δημοκρατική Αριστερά)	3	16
Δημοκρατικό Κοινωνικό Κίνημα	1	7
Δημοκρατικό Πατριωτικό Κίνημα ΝΙΚΗ	1	9
Ελληνική Λύση - Κυριάκος Βελόπουλος	3	9
Ένωση Κεντρώων	1	7
Κίνημα Αλλαγής	5	18
Κομμουνιστικό Κόμμα Ελλάδας	14	55

Λαϊκή Ενότητα	7	15
Λαϊκός Ορθόδοξος Συναγερμός	1	17
Λαϊκός Σύνδεσμος - Χρυσή Αυγή	3	21
Μέτωπο Ευρωπαϊκής Ρεαλιστικής Ανυπακοής (ΜΕΡΑ25)	5	4
Νέα Δημοκρατία	68	466
Οικολόγοι Εναλλακτικοί (Ομοσπονδία Οικολογικών και Εναλλακτικών Οργανώσεων)	3	0
Πανελλήνιο Σοσιαλιστικό Κίνημα	55	365
ΠΑΣΟΚ - Κίνημα Αλλαγής	8	25
Πλεύση Ελευθερίας	4	4
Πολιτική Άνοιξη	2	10
Σπαρτιάτες	0	12
Συνασπισμός Ριζοσπαστικής Αριστεράς	73	174
Συνασπισμός Ριζοσπαστικής Αριστεράς - Προοδευτική Συμμαχία	18	39
Συνασπισμός της Αριστεράς, των Κινημάτων και της Οικολογίας	6	34
ΣΥΡΙΖΑ - Προοδευτική Συμμαχία	18	39
Το Ποτάμι	4	14

Τέλος, ενδιαφέρον προκαλεί ο μέσος όρος χρόνων του κάθε λόγου. Συγκεκριμένα, στο διάγραμμα της εικόνας 4.7, αποτυπώνουμε τον ημερήσιο μέσο όρο που αφιερώνει κάθε βουλευτής στον λόγο του. Στις περισσότερες ημερομηνίες, φαίνεται ο χρόνος να κυμαίνεται στην περιοχή των 100 δευτερόλεπτων, δηλαδή περίπου στα δύο λεπτά. Κάτι τέτοιο δεν αποτελεί κανόνα, μιας και αρκετές είναι οι περιπτώσεις όπου αυτό το νούμερο ξεφεύγει. Να σημειώσουμε πως έχουμε οριοθετήσει τους παρακάτω χρόνους μεταξύ του 0 και του 1000, απομακρύνοντας έτσι ακραίες (λανθασμένες) τιμές.



Εικόνα 4.7 – Διάγραμμα μέσου χρόνου ομιλίας για όλες τις ημερομηνίες

4.2 Περίπτωση μελέτης στα δομημένα αρχεία – LDA Topic

Modeling

Μελετάμε τα δομημένα αρχεία, μορφής LegalDocML, προκειμένου να εξάγουμε θέματα σε μερίδα κείμενων, αξιοποιώντας τις δυνατότητες της τεχνολογίας Latent Dirichlet Allocation (LDA). Επιλέξαμε να αντληθεί το κείμενο των διαλόγων αποκλειστικά από τα αρχεία LegalDocML, μιας και από αυτά τα δομημένα έγγραφα η διαδικασία εξαγωγής του κειμένου είναι πια απλή και πιο ασφαλής.

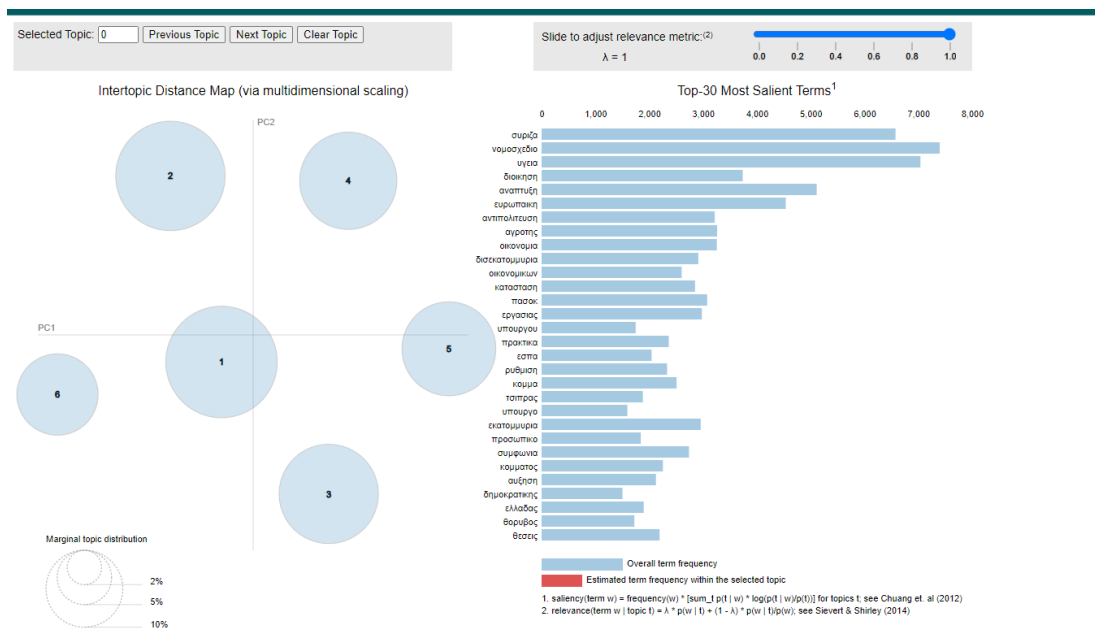
Ο στόχος χρήσης της LDA για την ανάλυση των κυβερνητικών συζητήσεων είναι να μοντελοποιήσει τα θέματα (topic modelling). Πιο συγκεκριμένα, επιδιώκει να συμπυκνώσει συνομιλίες ετών σε διαχειρίσιμα θέματα, ρίχνοντας φως στα επαναλαμβανόμενα θέματα και ζητήματα. Το κύριο σημείο αυτού του υπο-συστήματος είναι ότι ταξινομεί το κείμενο και βρίσκει λέξεις και φράσεις που συνυπάρχουν τακτικά. Οι κοινοί όροι κατηγοριοποιούνται σε θέματα, καθένα από τα οποία αντιπροσωπεύει ένα διαφορετικό ζήτημα ή τομέα μελέτης. Αυτά τα θέματα θα μπορούσαν να περιλαμβάνουν τα πάντα, από την εξωτερική πολιτική και τις οικονομικές μεταρρυθμίσεις μέχρι την υγειονομική περίθαλψη και την εκπαίδευση στις κυβερνητικές συζητήσεις.

Πριν την μέθοδο LDA είναι σύνηθες να γίνεται μια προεργασία στο εκάστοτε κείμενο. Αυτή συνήθως αποτελείται από λημματοποίηση (lemmatization) και stemming κάθε λέξης. Στην παρούσα εργασία, εκτελέσαμε μόνο την πρώτη διαδικασία καθώς η ιδιαιτερότητα της ελληνικής γλώσσας δυσκόλεψε διπλά το έργο μας, μιας και υπάρχουν πολλοί μορφικοί και συντακτικοί κανόνες που διέπουν τις λέξεις. Να επισημάνουμε πως στην παρούσα εργασία επιλέξαμε το κάθε σύνολο κειμένων να περιλαμβάνει αρχεία ανά έτος.

Παρόλο που η LDA χρησιμοποιεί εξελεγμένες μεθόδους, εξακολουθούν να υπάρχουν ζητήματα όπως ο θόρυβος των δεδομένων κειμένου και η ερμηνευσιμότητα των θεμάτων. Προκειμένου να ξεπεραστεί η πρώτη δυσκολία του θορύβου είναι πάγια τακτική η απομάκρυνση κάποιων επαναλαμβανόμενων λέξεων – stopwords – που δεν προσφέρουν πραγματικό νόημα στο τελικό αποτέλεσμα. Τέτοιες φράσεις είναι συνήθως αντωνυμίες, ρήματα ή/και λέξεις παραπλήσιες στον θεματικό πυλώνα που μελετάμε.

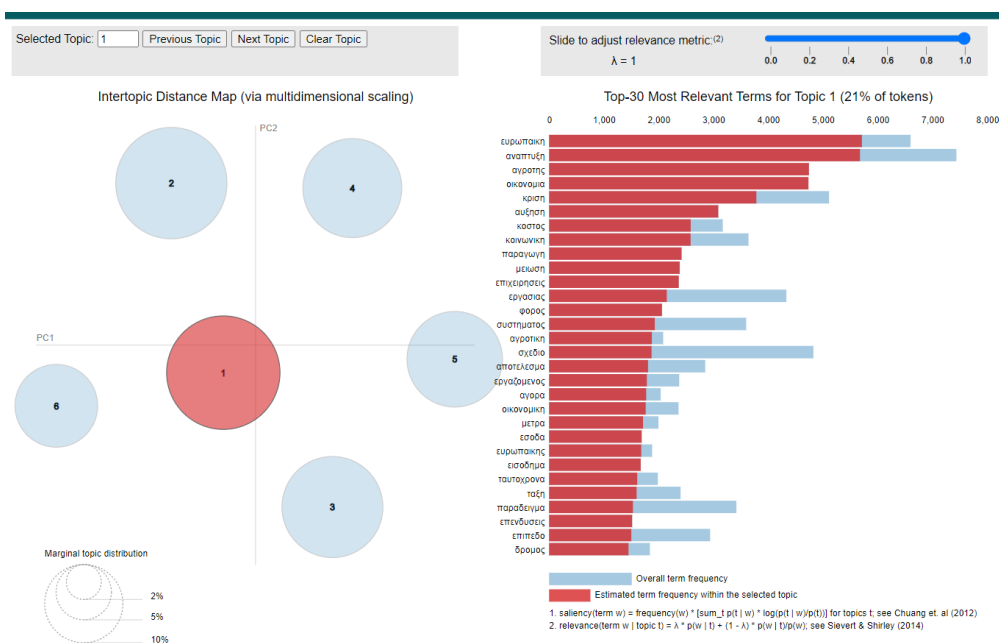
Στην μεθοδολογία μας, το αποτέλεσμα της μεθόδου LDA το αξιοποιούμε καταλλήλως όπου μέσω της βιβλιοθήκης της Python, “*pyldavis*”, οπτικοποιούμε και αποτυπώνουμε όπως φαίνεται στην εικόνα 4.8. Στο πρακτικό επίπεδο, στο δεξί μέρος φαίνεται μία λίστα με τους τριάντα (30) πιο «σημαντικούς» όρους, με την αντίστοιχη συχνότητα εμφάνισης. Στα αριστερά, φαίνεται ένα διάγραμμα με τα θέματα που έχουν προκύψει και την μεταξύ τους συσχέτιση ανάλογα με την απόστασή τους. Τέλος, με την επιλογή ενός εκ των θεμάτων, προσαρμόζεται καταλλήλως η λίστα και η συχνότητα των λέξεων.

Ενδεικτικά να αναφέρουμε πως στην παρακάτω εικόνα φαίνονται τα αποτελέσματα των πρακτικών όλου του έτους 2016. Οι κύριοι όροι για αυτό το έτος που φαίνεται από το ιστόγραμμα να υπερέχουν είναι ‘νομοσχέδιο’, ‘υγεία’, ‘σύριζα’, ‘ανάπτυξη’ και ‘ευρωπαϊκή’.



Εικόνα 4.8 – Κύρια οθόνη Topic Modeling αποτελέσματος με LDA για έτος 2016

Εστιάζοντας παραπάνω για το έτος 2016, προσπαθούμε να υποθέσουμε μερικά από τα θέματα που συζητήθηκαν στο Ελληνικό Κοινοβούλιο. Τυχαιά επιλέγουμε το θέμα νούμερο 1 (topic 1) και παρατηρούμε πως στην κορυφή βρίσκονται όροι που αφορούν την οικονομία, την Ευρωπαϊκή Ένωση, μια κρίση, αλλά και παράγοντες όπως η εργασία και η αγροτική ζωή (εικόνα 4.9). Συνδέοντας όλα αυτά, ένα πιθανό θέμα που προκύπτει αφορά μια οικονομική κρίση (ή/και ανάπτυξη) υπό την αιγίδα του ευρωπαϊκού κόλπου, πάνω σε τομείς όπως η εργασία και παραγωγική κτηνοτροφία. Αυτό το θέμα προφανώς αποτελεί μια υπόθεση, η οποία μπορεί εύκολα να επιβεβαιωθεί ή να διαψευστεί μέσω μιας απλής αναζήτησης και έρευνας στο διαδίκτυο.



Εικόνα 4.9 – Οθόνη Topic Modeling αποτελέσματος με LDA για έτος 2016 με επιλεγμένο το θέμα 1

Αναλύοντας και επεξεργάζοντας περαιτέρω τα αποτελέσματα που προέκυψαν από το υποσύστημά μας, συγκεντρώσαμε σε έναν ενιαίο πίνακα (ΠΙΝΑΚΑΣ 4.3) τους δέκα πιο σημαντικούς όρους των συζητήσεων συνολικά για κάθε έτος. Αυτοί οι όροι προσδιορίστηκαν με βάση τη σημασία και τη συνάφειά τους στο σύνολο δεδομένων, αναζητώντας κάθε φορά την συχνότητα εμφάνισής τους. Με αυτόν τον τρόπο, μπορούμε να εξάγουμε πληροφορίες για τα βασικά θέματα και τις τάσεις που επικρατούν στα δεδομένα που αναλύθηκαν.

Με μια πιο προσεκτική μελέτη του πίνακα εξάγουμε κάποια αναμενόμενα και κάποια μη αναμενόμενα συμπεράσματα. Αρχικά, παρατηρούμε πως η Ευρώπη βρίσκεται σταθερά στην κορυφή των σημαντικότερων όρων από τα τέλη της δεκαετίας του 1990, από όπου εντάσσεται η Ελλάδα στους κόλπους της, υπογραμμίζοντας την σημασία αυτού του γεγονότος για την πολιτική σκηνή της χώρας. Στην συνέχεια, οι φράσεις που άντεξαν στο χρόνο και συζητιούνται μέχρι και σήμερα σχετίζονται με την "ανάπτυξη", την "υγεία", την "εργασία" και το "νομοσχέδιο". Το γεγονός ότι εξακολουθούν να αποτελούν καυτά θέματα αποτελεί αποδεικτικό στοιχείο για την σημασία που δείχνουν τα πολιτικά πρόσωπα του κοινοβουλίου για αυτά τα ζητήματα που αφορούν τον γενικό πληθυσμό. Επιπλέον, είναι ιδιαίτερα ενδιαφέρον το πώς οι όροι 'οικονομική' και 'κρίση' άρχισαν να χρησιμοποιούνται έντονα τα τελευταία περίπου 15 χρόνια, με αφορμή τις οικονομικές ανησυχίες που διέπουν την χώρα από το 2008 και έπειτα.

ΠΙΝΑΚΑΣ 4.3 – Πρώτοι δέκα σημαντικοί όροι για κάθε έτος (από μοντέλο LDA)

Top 10 Most Salient Terms Overall (per year)

Year:1989	Year:1990	Year:1991	Year:1992
πασοκ: 2806	πασοκ: 7641	νομοσχεδιο: 6247	νομοσχεδιο: 7216
νομοσχεδιο: 1073	νομοσχεδιο: 5541	πασοκ: 5008	κοινοτητα: 6299
συμβουλιο: 855	αυξηση: 3615	κοινοτητα: 4410	υγεια: 5775
δικη: 809	κοινοτητα: 3602	αθηνα: 4176	πασοκ: 4733
δικαιωμα: 794	αθηνα: 3288	διαταξη: 3488	διαταξη: 3646
κοινοτητα: 728	οικονομια: 3255	δικη: 3233	γεωργιας: 3496
συμβουλίου: 713	δικαιωμα: 3206	διοικηση: 3034	αναπτυξη: 3345
διαταξη: 694	διαταξη: 2840	δικαιωμα: 3020	δικαιωμα: 3316
επιτροπης: 690	οικονομικων: 2840	βουλευτων: 2833	αυξηση: 3301
πλευρα: 671	αναπτυξη: 2812	παιδεια: 2820	διοικηση: 3211
Year:1993	Year:1994	Year:1996	Year:1997
νομοσχεδιο: 5428	νομοσχεδιο: 4896	αναπτυξη: 1551	νομοσχεδιο: 8581
πασοκ: 4611	κοινοτητα: 4104	ευρωπαϊκη: 1173	υγεια: 5904
κοινοτητα: 3893	πασοκ: 3582	πασοκ: 1146	αναπτυξη: 5593
διαταξη: 3649	υγεια: 2618	νομοσχεδιο: 1037	κοινοτητα: 5090
δικαιωμα: 2798	διαταξη: 2481	αυξηση: 940	δημος: 4588
δικη: 2454	αναπτυξη: 2308	εθνικη: 929	διοικηση: 4442
υγεια: 2329	ρυθμιση: 2222	δραχμας: 880	διαταξη: 4094
συμβουλιο: 2314	δικη: 2125	οικονομικων: 860	οικονομικων: 4015
εθνικης: 2283	εθνικη: 2046	εθνικης: 841	γεωργιας: 4006
εθνικη: 2234	μερος: 2033	διαταξη: 827	μερος: 3975
Year:1998	Year:1999	Year:2000	Year:2001
νομοσχεδιο: 2580	νομοσχεδιο: 5282	νομοσχεδιο: 4082	νομοσχεδιο: 6970
αναπτυξη: 2369	αναπτυξη: 4413	αναπτυξη: 2941	υγεια: 5021
υγεια: 2264	υγεια: 3897	πασοκ: 2541	αναπτυξη: 4437
πασοκ: 2080	ευρωπαϊκη: 3699	διοικηση: 2385	ευρωπαϊκη: 3965
παιδεια: 2041	συμβαση: 3399	ευρωπαϊκη: 2153	πασοκ: 3496
δημοσιас: 2030	παιδεια: 3298	στηριξη: 1814	διαταξη: 3149
δραχμας: 2018	διοικηση: 2875	εργασιας: 1797	διοικηση: 2907
ευρωπαϊκη: 1986	ρυθμιση: 2873	οικονομια: 1793	ελεγχος: 2669
εργασιας: 1964	πασοκ: 2857	εκπαιδευση: 1785	γεωργιας: 2596
διοικηση: 1909	δραχμας: 2748	ρυθμιση: 1753	ρυθμιση: 2541
Year:2002	Year:2003	Year:2004	Year:2005
νομοσχεδιο: 3848	νομοσχεδιο: 6007	αναπτυξη: 5524	πασοκ: 9822
αναπτυξη: 2707	υγεια: 4812	νομοσχεδιο: 5186	νομοσχεδιο: 9550
ευρωπαϊκη: 2641	αναπτυξη: 4128	πασοκ: 4787	αναπτυξη: 8376
διοικηση: 2421	ευρωπαϊκη: 3572	υγεια: 3437	ευρωπαϊκη: 6355
πασοκ: 2290	δημοσιас: 3358	διοικηση: 3102	υγεια: 5296

υγεια: 1970	δημος: 3357	ευρωπαϊκη: 2815	αντιπολιτευση: 3858
ελεγχος: 1854	διοικηση: 2877	αντιπολιτευση: 2397	οικονομια: 3545
διαταξη: 1831	εργασιας: 2833	ρυθμιση: 2357	αγορα: 3360
δικη: 1724	ελεγχος: 2477	οικονομια: 2146	εκπαιδευση: 3352
εργασιας: 1716	πασοκ: 2467	απασχοληση: 2129	σχεδιο: 3321
Year:2006	Year:2007	Year:2008	Year:2009
πασοκ: 8896	πασοκ: 8668	πασοκ: 11661	πασοκ: 8367
αναπτυξη: 6814	αναπτυξη: 5881	νομοσχεδιο: 9415	νομοσχεδιο: 4925
νομοσχεδιο: 6667	νομοσχεδιο: 5648	αναπτυξη: 7828	αναπτυξη: 4889
υγεια: 3629	υγεια: 5432	ευρωπαϊκη: 5575	κριση: 4889
ευρωπαϊκη: 3573	ευρωπαϊκη: 3142	υγεια: 5383	υγεια: 3561
εκπαιδευση: 3351	αντιπολιτευση: 3123	κοινωνια: 4836	οικονομια: 3166
αυξηση: 3081	παιδεια: 2826	αντιπολιτευση: 4460	ευρωπαϊκη: 3089
αντιπολιτευση: 3020	αυξηση: 2654	κριση: 4379	κατασταση: 2792
δημος: 2904	σχεδιο: 2562	σχεδιο: 4317	εργασιας: 2620
σχεδιο: 2728	διοικηση: 2512	εκπαιδευση: 4146	σχεδιο: 2615
Year:2010	Year:2011	Year:2012	Year:2013
νομοσχεδιο: 10264	νομοσχεδιο: 8331	αναπτυξη: 4997	νομοσχεδιο: 7828
πασοκ: 9184	πασοκ: 8141	νομοσχεδιο: 4679	υγεια: 5840
αναπτυξη: 6681	αναπτυξη: 6913	υγεια: 4262	αναπτυξη: 5752
υγεια: 5513	υγεια: 6310	κριση: 3376	συριζα: 4552
κριση: 5161	κριση: 5178	πασοκ: 3210	εργασιας: 4422
οικονομια: 4182	ευρωπαϊκη: 4501	ευρωπαϊκη: 2910	οικονομικων: 4060
ευρωπαϊκη: 4047	κοινωνια: 4197	οικονομικων: 2844	κριση: 3849
κοινωνια: 3883	οικονομια: 4101	εργασιας: 2624	σχεδιο: 3814
σχεδιο: 3849	κατασταση: 3960	κατασταση: 2512	ευρωπαϊκη: 3584
κατασταση: 3822	σχεδιο: 3439	κοινωνια: 2457	κοινωνια: 3159
Year:2014	Year:2015	Year:2016	Year:2017
νομοσχεδιο: 8100	νομοσχεδιο: 5220	νομοσχεδιο: 6650	νομοσχεδιο: 5728
υγεια: 6462	συριζα: 4416	συριζα: 5927	υγεια: 5147
αναπτυξη: 5841	ευρωπαϊκη: 2645	αναπτυξη: 5261	συριζα: 4125
συριζα: 4288	υγεια: 2419	υγεια: 4621	αναπτυξη: 3935
ευρωπαϊκη: 3568	κριση: 2375	οικονομια: 3414	εργασιας: 2942
εργασιας: 3364	αναπτυξη: 2257	ευρωπαϊκη: 3375	πασοκ: 2523
κριση: 3225	συμφωνια: 2218	κριση: 2963	αντιπολιτευση: 2488
ρυθμιση: 2961	οικονομια: 2210	πασοκ: 2915	ευρωπαϊκη: 2440
ενεργεια: 2888	κοινωνια: 2041	σχεδιο: 2894	οικονομια: 2172
σχεδιο: 2852	παιδεια: 2027	αντιπολιτευση: 2894	ρυθμιση: 2107

Year:2018	Year:2019	Year:2020	Year:2021
συριζα: 13111	συριζα: 20572	συριζα: 25738	συριζα: 26295
νδ: 5409	νομοσχεδιο: 5735	υγεια: 10390	νομοσχεδιο: 10080
νομοσχεδιο: 4535	αναπτυξη: 4520	αλλαγης: 9414	υγεια: 9285
αναπτυξη: 3690	αθηνων: 4469	νομοσχεδιο: 9395	αλλαγης: 8982
υγεια: 2828	υγεια: 4358	αθηνων: 8489	κινημα: 7746
κεντρων: 2445	αλλαγης: 3608	κινημα: 8174	λυση: 7722
εργασιας: 2407	νδ: 3557	λυση: 8122	αναπτυξη: 6670
συμφωνια: 2382	συμφωνια: 3537	κκε: 6459	κκε: 6034
σχεδιο: 2313	κκε: 3457	αναπτυξη: 5596	εργασιας: 5950
ευρωπαϊκη: 2308	κωνσταντινος: 3371	κριση: 5023	αθηνων: 5296
Year:2022	Year:2023		
συριζα: 21348	συριζα: 6924		
υγεια: 9380	υγεια: 5362		
αλλαγης: 8548	νομοσχεδιο: 5126		
νομοσχεδιο: 8436	λυση: 4047		
αθηνων: 7826	αλλαγης: 3255		
λυση: 7369	κκε: 2973		
κινημα: 7118	πασοκ: 2957		
ενεργεια: 6568	κινημα: 2712		
αναπτυξη: 5950	αναπτυξη: 2599		
κκε: 5727	εργασιας: 2413		

4.3 Στατιστικά Δεδομένα

Σε αυτή την παράγραφο της διπλωματικής εργασίας παραθέτουμε κάποιους πίνακες στους οποίους αποτυπώνονται στατιστικά στοιχεία, αντλημένα από τα δομημένα κοινοβουλευτικά αρχεία.

Στον πρώτο πίνακα, ΠΙΝΑΚΑΣ 4.4, καταγράφουμε στατιστικά για το πλήθος των αρχικών αρχείων της βάσης δεδομένων που μετατρέψαμε σε δομημένα αρχεία xml, μορφής LegalDocML, χωρισμένα ανά έτος. Ιδιαίτερης σημασίας είναι το γεγονός ότι σχεδόν σε όλες τις περιπτώσεις το ποσοστό επιτυχίας ξεπερνάει το φράγμα του 90%, ακόμα και το συνολικό ποσοστό για όλα τα έτη αθροιστικά, με εξαίρεση το έτος 1995 που δεν έχουμε αρχεία.

ΠΙΝΑΚΑΣ 4.4 – Ετήσια στατιστικά στοιχεία μετατροπής αρχικών αρχείων σε XML

Έτος	#εγιναν xml	#αποτυχίας	Ποσοστό επιτυχίας %
1989	45	3	94
1990	162	11	94
1991	169	14	92
1992	178	27	87
1993	150	11	93
1994	131	3	98
1995	0	0	-
1996	46	4	92

1997	191	5	97
1998	90	4	96
1999	170	10	94
2000	143	22	87
2001	194	15	93
2002	150	17	90
2003	165	9	95
2004	141	8	95
2005	194	14	93
2006	186	2	99
2007	180	8	96
2008	209	7	97
2009	148	2	99
2010	210	2	99
2011	231	1	100
2012	173	0	100
2013	206	1	100
2014	175	7	96
2015	140	4	97
2016	200	2	99
2017	182	3	98
2018	171	2	99
2019	155	3	98
2020	196	5	98
2021	192	1	99
2022	194	2	99
2023	104	1	99
ΣΥΝΟΛΙΚΑ	5.571	230	96

Στον ΠΙΝΑΚΑΣ 4.5, καταγράφουμε παρόμοια στατιστικά με την διαφορά ότι σε αυτήν την περίπτωση έχουμε κάνει διαχωρισμό με βάση την κοινοβουλευτική περίοδο. Ομοίως τα στοιχεία είναι ιδιαίτερα ικανοποιητικά για σχεδόν όλες τις περιόδους, με το ποσοστό επιτυχίας να κυμαίνεται σε πολλές περιπτώσεις άνω του 95%.

ΠΙΝΑΚΑΣ 4.5 – Στατιστικά μετατροπής XML ανά κοινοβουλευτική περίοδο

Αριθμός Περιόδου*	#εγιναν xml	#αποτυχίας	Ποσοστό επιτυχίας %
Ε' ΠΕΡΙΟΔΟΣ	33	2	94
Ζ' ΠΕΡΙΟΔΟΣ	592	57	91
Η' ΠΕΡΙΟΔΟΣ	163	6	96
Θ' ΠΕΡΙΟΔΟΣ	537	29	95
Ι' ΠΕΡΙΟΔΟΣ	634	60	91
ΙΑ' ΠΕΡΙΟΔΟΣ	627	24	96
ΙΒ' ΠΕΡΙΟΔΟΣ	372	13	97
ΙΓ' ΠΕΡΙΟΔΟΣ	552	4	99

ΙΑ΄ ΠΕΡΙΟΔΟΣ	3	0	100
ΙΕ΄ ΠΕΡΙΟΔΟΣ	478	8	98
ΙΖ΄ ΠΕΡΙΟΔΟΣ	685	9	99
ΙΗ΄ ΠΕΡΙΟΔΟΣ	705	10	99
ΙΘ΄ ΠΕΡΙΟΔΟΣ	3	0	100
ΙΣΤ΄ ΠΕΡΙΟΔΟΣ	92	3	97
Κ΄ ΠΕΡΙΟΔΟΣ	48	1	98
ΣΤ' ΠΕΡΙΟΔΟΣ	47	4	92
ΣΥΝΟΛΙΚΑ	5.571	230	96

* (ΠΡΟΕΔΡΕΥΟΜΕΝΗΣ ΚΟΙΝΟΒΟΥΛΕΥΤΙΚΗΣ ΔΗΜΟΚΡΑΤΙΑΣ)

5

Επίλογος

Στην τελευταία αυτή ενότητα, συνοψίζουμε παρουσιάζοντας συμπεράσματα σχετικά με το σύνολο της εργασίας και κάποια από τα βασικά προβλήματα που αντιμετωπίσαμε. Τέλος, προτείνουμε πιθανές μελλοντικές επεκτάσεις.

5.1 Σύνοψη και συμπεράσματα

Τα κοινοβουλευτικά πρακτικά παρέχονται με τρόπο που δεν ευνοούν την αποτελεσματική αναζήτηση και σημασιολογική ανάλυση. Συνήθως, το νομοθετικό υλικό παρουσιάζεται με τρόπο που καθιστά δύσκολη την εξαγωγή και χρήση δεδομένων για ολοκληρωμένη μελέτη. Ο περιορισμός αυτός απορρέει από την έλλειψη ενός ενιαίου και δομημένου πλαισίου που διευκολύνει τη σημασιολογική κατανόηση. Ως εκ τούτου, οι δυνατότητες της σημασιολογικής ανάλυσης των κοινοβουλευτικών διαδικασιών και των εξελιγμένων χαρακτηριστικών αναζήτησης παραμένουν ως επί το πλείστον αναξιοποίητες. Γίνεται όλο και πιο σαφές ότι απαιτείται μια δομημένη αναπαράσταση των κοινοβουλευτικών δεδομένων που θα υποστηρίζουν προηγμένες αναζητήσεις και ουσιαστική ανάλυση.

Ο κύριος στόχος της εργασίας είναι η μετατροπή των κοινοβουλευτικών συζητήσεων σε δομημένες μορφές που μπορούν να προσπελαστούν και να χρησιμοποιηθούν από άλλα συστήματα. Για την πραγμάτωση του στόχου αυτού φτιάξαμε ένα σύστημα που συνδυάζει στοιχεία όπως μια εξειδικευμένη γλώσσα DSL και πολλαπλών αναλυτών/parser. Αυτό συνιστά μια πρόοδο στην ενίσχυση της διασύνδεσης, της αξιοποίησης και της προσβασιμότητας των νομοθετικών διαδικασιών.

Η μετατροπή των πρακτικών του Ελληνικού Κοινοβουλίου σε ανοικτά διασυνδεδεμένα δεδομένα δημιουργεί νέες ευκαιρίες για τους μελετητές, τους φορείς λήψης αποφάσεων και το ευρύ κοινό ώστε να αλληλεπιδράσουν με το εκτεταμένο νομοθετικό υλικό του Ελληνικού Κοινοβουλίου. Κάναμε δυνατή την ομαλή σύνδεση με άλλα σύνολα δεδομένων και εφαρμογές, δομώντας τις συζητήσεις σε δομημένα αρχεία σημασιολογικής μορφής.

Καθ' όλη τη διάρκεια της διπλωματικής εργασίας, αντιμετωπίσαμε μια σειρά από ενδιαφέρουσες προκλήσεις. Μια τέτοια πρόκληση αφορούσε τις διαφορετικές συμβάσεις ονομασίας που χρησιμοποιούνται στο σύνολο δεδομένων για τις συζητήσεις στο Ελληνικό Κοινοβούλιο. Για παράδειγμα, η συμπερίληψη λέξεων σε διαφορετικές μορφές όπως για παράδειγμα τόσο του "Αλέξη" όσο και του "Αλέξιου" Τσίπρα, καθώς και η εσφαλμένη χρήση

αγγλικών χαρακτήρων σε σημεία όπου ο ελληνικός με τον αγγλικό είναι όμοιος, αποτέλεσαν έναν ενδιαφέρον γρίφο στην αναγνώριση ονομάτων. Λόγω του γεγονότος ότι τα πρακτικά του Ελληνικού Κοινοβουλίου έχουν συγγραφεί από άνθρωπο, τέτοια ορθογραφικά λάθη είναι λογικό και αναμενόμενο να υπάρχουν. Αυτή η περιπλοκότητα, αν και δεν ήταν η πρωταρχική εστίαση αυτής της εργασίας, ανέδειξε την πολυπλοκότητα που υπάρχει στην διαδικασία της ακριβούς ταυτοποίησης και αναγνώρισης ονομάτων σε ένα τέτοιο μεγάλο σύνολο δεδομένων.

Ακόμα, επιπλέον πολυπλοκότητα εισήγαγε η εναλλαγή του τρόπου γραφής των νομοθετικών διαδικασιών σε διαφορετικές κοινοβουλευτικές περιόδους. Οι συγγραφείς επέλεξαν κατά διαστήματα να τροποποιήσουν τη δομή και το περιεχόμενο των εισαγωγικών αποσπασμάτων. Παρατηρήσαμε περιπτώσεις στις οποίες οι συγγραφείς επέλεξαν είτε να παραλείψουν εισαγωγικά στοιχεία, είτε να αλλάξουν την σειρά τους, είτε να υιοθετήσουν διαφορετικές συντακτικές προσεγγίσεις. Αυτή η ποικιλομορφία αποτέλεσε εμπόδιο για τη διασφάλιση της ομοιομορφίας και της συνέπειας στην ανάλυση των δεδομένων, καθιστώντας αναγκαία μια διαφοροποιημένη προσέγγιση για να ληφθούν υπόψη αυτές οι διακυμάνσεις στην παρουσίαση του κοινοβουλευτικού λόγου σε διάφορα χρονικά πλαίσια.

5.2 Μελλοντικές επεκτάσεις

Έχοντας ως σημείο εκκίνησης την παρούσα εργασία και τα συμπεράσματα που προέκυψαν από αυτήν, μπορούν να προταθούν διάφορες μελλοντικές επεκτάσεις για τη βελτίωση και την ενίσχυση των αποτελεσμάτων της.

Μια μελλοντική κατεύθυνση θα μπορούσε να περιλαμβάνει την ενσωμάτωση εργαλείων και τεχνικών οπτικοποίησης δεδομένων, γεγονός που θα μπορούσε να βελτιώσει σημαντικά την προσβασιμότητα και τη χρηστικότητα των μετατρεπόμενων κοινοβουλευτικών δεδομένων. Η δημιουργία διαδραστικών οπτικών αναπαραστάσεων, όπως γραφήματα, διαγράμματα και χρονοδιαγράμματα, θα μπορούσε να παρέχει στους χρήστες διαισθητικούς τρόπους για την εξερεύνηση και την ανάλυση των συζητήσεων. Αυτές οι οπτικοποιήσεις θα μπορούσαν να προσφέρουν πολύτιμες γνώσεις σχετικά με τα μοτίβα, τις τάσεις και τις σχέσεις εντός των νομοθετικών δεδομένων, διευκολύνοντας τους ερευνητές, τους υπεύθυνους χάραξης πολιτικής και το ευρύ κοινό να κατανοήσουν τις πολιτικές διαδικασίες.

Επιπλέον, η επέκταση του πεδίου εφαρμογής, ώστε να ενσωματωθεί η συναισθηματική ανάλυση των λόγων, θα μπορούσε να αποτελέσει μια δυναμική κατεύθυνση. Στην ανθρώπινη επικοινωνία, τα συναισθήματα είναι ζωτικής σημασίας επειδή επηρεάζουν την σκέψη, τον λόγο και τις αποφάσεις μας. Η ενσωμάτωση αλγορίθμων ανάλυσης συναισθημάτων στο τρέχον σύνολο νομοθετικών πληροφοριών μπορεί να προσφέρει μια εικόνα των κινήτρων πίσω από τα νομοθετικά επιχειρήματα. Η προσέγγιση αυτή μπορεί να διευκολύνει την εξαγωγή του συναισθηματικού τόνου από τα δομημένα αρχεία, παρέχοντας μια βαθύτερη κατανόηση των συνομιλιών. Η αναγνώριση του συναισθηματικού υπόβαθρου των κοινοβουλευτικών διαδικασιών μπορεί να βοηθήσει τους μελετητές να κατανοήσουν καλύτερα τα επιχειρήματα και να παρέχουν μια πιο ολοκληρωμένη εικόνα του πολιτικού περιβάλλοντος.

Ένας άλλος εναλλακτικός τρόπος επέκτασης θα μπορούσε να περιλαμβάνει τη διερεύνηση της ενσωμάτωσης συνδεδεμένων δεδομένων από πολλαπλές πηγές πέραν των κοινοβουλευτικών συζητήσεων. Η ενσωμάτωση δεδομένων από συναφείς τομείς, όπως για παράδειγμα αυτόν της ελληνικής νομοθεσίας, θα μπορούσε να δημιουργήσει ένα ολοκληρωμένο οικοσύστημα συνδεδεμένων δεδομένων της ελληνικής διακυβέρνησης. Αυτή η

διεπιστημονική προσέγγιση θα επέτρεπε στους πολίτες να διεξάγουν πιο ολοκληρωμένες αναλύσεις, αποκτώντας βαθύτερες γνώσεις για τις περίπλοκες σχέσεις μεταξύ των νομοθετικών αποφάσεων με τους διαφορετικούς τομείς.

Επιπλέον, ένα επόμενο βήμα περιλαμβάνει μια συνολική αξιολόγηση ολόκληρου του συστήματος με σκοπό την επιβεβαίωση του βέλτιστου τρόπου λειτουργίας του. Συγκεκριμένα, κύριος στόχος είναι να διασφαλιστεί η μετατροπή όλων των αρχικών αρχείων σε δομημένη μορφή. Αυτό συνεπάγεται την ανάγκη εφαρμογής διαδικασίας εντοπισμού και επίλυσης πιθανών σφαλμάτων, που προέρχονται από την συντακτική ιδιομορφία του περιεχομένου κάποιων αρχείων. Απαραίτητη θα είναι η ξεχωριστή εξέταση του κάθε αρχείου. Συνεπώς, με την αντιμετώπιση αυτών των ζητημάτων, θα ενισχυθεί η πληρότητα της συνολικής διαδικασίας μετατροπής.

Τέλος, λαμβάνοντας υπόψη τις ραγδαίες εξελίξεις στην τεχνητή νοημοσύνη και τη μηχανική μάθηση, η διερεύνηση της εφαρμογής αυτών των τεχνολογιών στην ανάλυση των κοινοβουλευτικών συζητήσεων θα μπορούσε να αποτελέσει μια διαφορετική προσέγγιση. Η ανάπτυξη μοντέλων τεχνητής νοημοσύνης για την εξαγωγή πολύτιμων πληροφοριών, την πρόβλεψη νομοθετικών τάσεων ή ακόμη και την αυτοματοποίηση ορισμένων πτυχών της πολιτικής ανάλυσης καθιστώντας τη νομοθετική διαδικασία πιο αποτελεσματική.

Αυτές οι πιθανές επεκτάσεις, οι οποίες βασίζονται στα θεμέλια που έθεσε η παρούσα διπλωματική εργασία, ανοίγουν τον δρόμο για καινοτόμες λύσεις που μπορούν να βελτιώσουν περαιτέρω την προσβασιμότητα, τη χρηστικότητα και τον αντίκτυπο των κοινοβουλευτικών δεδομένων στην νέα ψηφιακή εποχή.

6

Βιβλιογραφία

- [1] Andersen, A. B., Gür, N., Hose, K., Jakobsen, K. A., & Pedersen, T. B. (2015). Publishing Danish agricultural government data as semantic web data. In *Lecture Notes in Computer Science* (pp. 178–186). https://doi.org/10.1007/978-3-319-15615-6_13
- [2] Antoniou, G., & VanHarmelen, F. (2004). A Semantic Web Primer. <https://dl.acm.org/doi/10.5555/975284>
- [3] Berners-Lee, T., Hendler, J. A., & Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 34–43. <https://doi.org/10.1038/scientificamerican0501-34>
- [4] Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22. <https://doi.org/10.4018/jswis.2009081901>
- [5] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. <https://doi.org/10.1145/2133806.2133826>
- [6] Blei, D. M., Ng, A. Y., & Jordan, M. (2001). Latent dirichlet allocation. *ResearchGate*. https://www.researchgate.net/publication/221620547_Latent_Dirichlet_Allocation
- [7] Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M., & Horrocks, I. (2000). The Semantic Web: the roles of XML and RDF. *IEEE Internet Computing*, 4(5), 63–73. <https://doi.org/10.1109/4236.877487>

- [8] Ding, L., Peristeras, V., & Hausenblas, M. (2012). Linked Open Government Data [Guest editors' introduction]. *IEEE Intelligent Systems*, 27(3), 11–15. <https://doi.org/10.1109/mis.2012.56>
- [9] Ding, L., Zhou, L., Finin, T., & Joshi, A. (2005). *How the Semantic Web is Being Used: An Analysis of FOAF Documents*. <https://doi.org/10.1109/hicss.2005.299>
- [10] Dritsa, K., Thoma, K., Pavlopoulos, J., & Louridas, P. (2022). *A Greek Parliament Proceedings dataset for computational linguistics and political analysis*. <https://doi.org/10.48550/arxiv.2210.12883>
- [11] DuCharme, B. (2013). *Learning SPARQL: Querying and Updating with SPARQL 1.1*. <https://search.worldcat.org/title/1066605615>
- [12] Erjavec, T. (2022). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1), 415–448. <https://doi.org/10.1007/s10579-021-09574-0>
- [13] Erjavec, T., Kopp, M., & Meden, K. (2023). TEI and Git in ParlaMint: Collaborative Development of Language Resources. *Linköping Electronic Conference Proceedings*. <https://doi.org/10.3384/ecp198005>
- [14] Geiger, C., & Von Lucke, J. (2012). Open Government and (Linked) (Open) (Government) (Data). *eJournal of eDemocracy and Open Government*, 4(2), 265–278. <https://doi.org/10.29379/jedem.v4i2.143>
- [15] Ghose, A., Lissandrini, M., Hansen, E. R., & Weidema, B. P. (2021). A core ontology for modeling life cycle sustainability assessment on the Semantic Web. *Journal of Industrial Ecology*, 26(3), 731–747. <https://doi.org/10.1111/jiec.13220>
- [16] Goyvaerts, J. (2006). *Regular Expressions: The Complete Tutorial*. Lulu Press. <https://dl.acm.org/doi/book/10.5555/1205629>
- [17] Graves, M., Constabaris, A., & Brickley, D. (2007). FOAF: Connecting People on the Semantic Web. *Cataloging & Classification Quarterly*, 43(3–4), 191–202. https://doi.org/10.1300/j104v43n03_10
- [18] Hitzler, P., Krötzsch, M., & Rudolph, S. (2009). *Foundations of Semantic Web Technologies*. In *Chapman and Hall/CRC eBooks*. <https://doi.org/10.1201/9781420090512>
- [19] Jelodar, H., Wang, Y., Yuan, C., Xia, F., Jiang, X., Li, Y., & Zhao, L. (2018). Latent Dirichlet allocation (LDA) and topic modeling: models, applications,

- a survey. *Multimedia Tools and Applications*, 78(11), 15169–15211.
<https://doi.org/10.1007/s11042-018-6894-4>
- [20] Palmirani, M., & Vitali, F. (2011). Akoma-Ntoso For legal documents. In *Law, governance and technology series* (pp. 75–100).
https://doi.org/10.1007/978-94-007-1887-6_6
- [21] Parr, T. (2013). *The Definitive ANTLR 4 reference*. Pragmatic Bookshelf.
<https://dl.acm.org/doi/10.5555/2501720>
- [22] Romary, L., & Hudrisier, H. (2002). TEI – Text encoding initiative. *Études Et Documents Berbères*, N° 19-20(1), 319–322.
<https://doi.org/10.3917/edb.019.0319>
- [23] Tummarello, G., Delbru, R., & Oren, E. (2007). Sindice.com: Weaving the open linked data. In *Lecture Notes in Computer Science* (pp. 552–565).
https://doi.org/10.1007/978-3-540-76298-0_40
- [24] Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2016). The debates of the European Parliament as Linked Open Data. *Semantic Web*, 8(2), 271–281. <https://doi.org/10.3233/sw-160227>
- [25] Van Veen, T. (2019). Wikidata: From “an” Identifier to “the” Identifier. *Information Technology and Libraries*, 38(2), 72–81.
<https://doi.org/10.6017/ital.v38i2.10886>
- [26] Πρακτικό Συνεδρίασης Ολομέλειας του Ελληνικού Κοινοβουλίου - Παρασκευή 8 Ιουνίου 2018.
https://www.hellenicparliament.gr/UserFiles/a08fc2dd-61a9-4a83-b09a-09f4c564609d/es20180608_1.pdf (Ημερομηνία πρόσβασης: 01 Νοεμβρίου 2023)