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

Οι Συζητήσεις Του Κοινοβουλίου Ως Διασυνδεδεμένα Ανοιχτά Δεδομένα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΠΑΠΑΝΙΚΟΛΑΟΥ ΙΩΑΝΝΗ

Επιβλέπων : Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.
Συνεπιβλέπων : Μάριος Κόνιαρης
Ε.ΔΙ.Π. ΕΜΠ

Αθήνα, Νοέμβριος 2023

Η σελίδα αυτή είναι σκόπιμα λευκή.

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

Οι Συζητήσεις Του Κοινοβουλίου Ως Διασυνδεδεμένα Ανοιχτά 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 1η Νοέμβριου 2023.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Παναγιώτης Τσανάκας	Βασιλική Καντερέ	Ευγενία Τζαννίνη
Καθηγητής Ε.Μ.Π.	Επίκουρη Καθηγήτρια Ε.Μ.Π.	Καθηγήτρια Ε.Μ.Π.

Αθήνα, Νοέμβριος 2023

(Υπογραφή)

.....

ΙΩΑΝΝΗΣ ΠΑΠΑΝΙΚΟΛΑΟΥ

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

Περίληψη

Η αυξανόμενη διαθεσιμότητα ψηφιακών δεδομένων και οι εξελίξεις στις τεχνολογίες σημασιολογικού ιστού έχουν ανοίξει νέες δυνατότητες για την ανάλυση και την εξαγωγή γνώσης από διάφορες πηγές. Η παρούσα διπλωματική εργασία διερευνά την εφαρμογή αυτών των τεχνολογιών για τη μετατροπή των συζητήσεων του Ελληνικού κοινοβουλίου σε συνδεδεμένα δεδομένα, επιτρέποντας πλουσιότερη και πιο ουσιαστική ανάλυση του νομοθετικού λόγου.

Η εργασία ξεκινά με την παρουσία ενός εκτενούς θεωρητικού υποβάθρου. Έπειτα εστιαζόμενο σε πιο πρακτικά ζητήματα, αναλύεται ένα σύστημα το οποίο ξενικά με την μετατροπή των ακατέργαστων κειμενικών δεδομένων των συζητήσεων του Ελληνικού Κοινοβουλίου σε δομημένα έγγραφα XML, τα οποία χρησιμεύουν ως ενδιάμεση αναπαράσταση, επιτρέποντας την περαιτέρω επεξεργασία και ανάλυση. Στη συνέχεια, τα έγγραφα XML μετατρέπονται σε μορφή σχήματος Resource Description Framework Schema (RDFS), επιτρέποντας την αναπαράσταση εννοιών, σχέσεων και μεταδεδομένων που σχετίζονται με τις συζητήσεις.

Εν συνεχεία, στόχος είναι η ανάδειξη της δύναμης των σημασιολογικών ερωτημάτων με τη χρήση της SPARQL, μιας γλώσσας ερωτημάτων για την αναζήτηση δεδομένων RDF. Με τη μετατροπή των συζητήσεων του Ελληνικού Κοινοβουλίου σε διασυνδεδεμένα δεδομένα και την παροχή ενός σχήματος RDFS, καθίσταται δυνατή η εκτέλεση σύνθετων ερωτημάτων, η ανάκτηση ουσιαστικών πληροφοριών και η δημιουργία συνδέσεων μεταξύ διαφόρων πτυχών των συζητήσεων. Η προσέγγιση αυτή επιδεικνύει τις δυνατότητες των τεχνολογιών του σημασιολογικού ιστού στην ενίσχυση της προσβασιμότητας, της ευρεσιμότητας και της χρηστικότητας των κοινοβουλευτικών διαδικασιών, συμβάλλοντας στον ευρύτερο τομέα της ηλεκτρονικής δημοκρατίας και της διαφανούς διακυβέρνησης.

Λέξεις Κλειδιά: « Σημασιολογικός Ιστός, Συνδεδεμένα Δεδομένα, XML, RDF, SPARQL, Οι Συζητήσεις Του Ελληνικού Κοινοβουλίου »

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

The increasing availability of digital data and developments in semantic web technologies have opened up new possibilities for analysing and extracting knowledge from various sources. This thesis explores the application of these technologies to transform the debates of the Greek parliament into linked data, allowing for richer and more meaningful analysis of legislative discourse.

The paper begins with the presentation of an extensive theoretical background. Then, focusing on more practical issues, a system is analysed which is capable of converting the raw textual data of the Greek Parliament debates into structured XML documents, which serve as an intermediate representation, allowing further processing and analysis. The XML documents are then converted to a Resource Description Framework Schema (RDFS) format, allowing the representation of concepts, relationships and metadata associated with the discussions.

Next, the aim is to demonstrate the power of semantic queries using SPARQL, a query language for querying RDF data. By converting the debates of the Greek Parliament into interlinked data and providing an RDFS schema, it is possible to perform complex queries, retrieve meaningful information and create links between different aspects of the debates. This approach demonstrates the potential of semantic web technologies in enhancing the accessibility, findability and usability of parliamentary procedures, contributing to the broader field of e-democracy and transparent governance.

Keywords: «Semantic Web, Linked Data, XML, RDF, SPARQL, Greek Parliament Debates»

Η σελίδα αυτή είναι σκόπιμα λευκή.

Ευχαριστίες

……
μπλα
μπλα μπλα

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 17

1.2.1 Συνεισφορά *18*

1.3 Σχετικά Παραδείγματα του Εξωτερικού 18

1.4 Οργάνωση κειμένου 19

2 Θεωρητικό υπόβαθρο **20**

2.1 Σημασιολογικός Ιστός – Semantic Web 20

2.2 Ανοιχτά Διασυνδεδεμένα Δεδομένα 22

2.3 Τεχνολογίες Υποστήριξης Δεδομένων 25

2.3.1 Extensible Markup Language (XML) *26*

2.3.2 Resource Description Framework (RDF) *27*

2.3.3 Λεξιλόγια Σημασιολογικού Ιστού *29*

2.3.4 SPARQL *31*

2.4 Ανοικτά Κυβερνητικά Δεδομένα – Akoma Ntoso 31

2.5 TEI ParlaMint 33

3 Σύστημα Διαχείρισης Πρακτικών **35**

3.1 Αρχιτεκτονική Συστήματος 35

3.2 Βάση Δεδομένων 36

3.3 Δομή Πρακτικών Βουλής 38
3.4 Εξαγωγή και Ανάλυση Κειμένων 42
3.4.1 ANLTR4 – REGEX 43
3.5 Διαχείριση Δεδομένων 45
3.5.1 Μετατροπή Αρχείων Κειμένου σε XML αρχεία 45
3.5.2 Μετατροπή Αρχείων XML σε RDF αρχεία 48
3.5.3 Επεξεργασία RDF – SPARQL 53
3.6 Μετατροπή Αρχείων Akoma Ntoso σε ParlaMint Tei 54
3.7 Σχολιασμός περί Σφαλμάτων 55
4 Μελέτη Περιπτώσεων και Στατιστικά Δεδομένα 57
4.1 Περίπτωση μελέτης RDF 57
4.2 Περίπτωση μελέτης XML – LDA Topic Modeling 59
4.3 Στατιστικά Δεδομένα 65
5 Επίλογος 68
5.1 Σύνοψη και συμπεράσματα 68
5.2 Εμπόδια – Προβλήματα 69
5.3 Μελλοντικές επεκτάσεις 69
6 Βιβλιογραφία 71

Πίνακας Εικόνων

Είκονα 2.1.1 – Η εξέλιξη του Διαδικτύου 1

Εικόνα 2.2.1 – The Linked Open Data Cloud 1

Εικόνα 2.2.2 – Ανάπτυξη των συνδεδεμένων ανοικτών δεδομένων από το 2007 1

Εικόνα 2.3.1 – Δείγμα Κώδικα XML 1

Εικόνα 2.3.2.1 – Βασική σύνταξη RDF ..... **Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.1.1 – Διάγραμμα Αρχιτεκτονικής Συστήματος ..... **Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.2.1 – Επίσημος ιστότοπος Ελληνικής Κυβέρνησης με Συνεδριάσεις Ολομέλειας**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.3.1 – Ενδεικτικό απόσπασμα πρακτικών – «ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ » (08-06-2018) **Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.3.2 – Ενδεικτικό απόσπασμα πρακτικών – «Θεμάτων» (08-06-2018) **Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.3.3 – Ενδεικτικό απόσπασμα πρακτικών – «Προεδρεύοντες» και «Ομιλητές» (08-06-2018) **Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.3.4 – Ενδεικτικό απόσπασμα πρακτικών – «Πρακτικά Βουλής» (08-06-2018)**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.3.5 – Ενδεικτικό απόσπασμα πρακτικών – Εισαγωγικός Πρόλογος (08-06-2018)**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.3.6 – Ενδεικτικό απόσπασμα πρακτικών – Μέρος Διαλόγων(08-06-2018) **Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.4.1.1 – Ενδεικτικό δέντρο που προκύπτει από την γραμματική ANLTR4**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.4.1.2 – Απόσπασμα από το συντακτικό της γραμματική**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.5.1.1 – Απόσπασμα από τα μεταδεδομένα ενός αρχείου xml-akoma ntoso**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.5.1.2 – Απόσπασμα από κύριο μέρος (debateBody) ενός αρχείου xml-akoma ntoso .. **Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.5.1.3 – Διάγραμμα για το στοιχείο “debate” ..... **Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.5.2.1 (α)-(δ) – Μερικές τριπλέτες RDF όπως φαίνονται σε ένα rdf/xml αρχείο**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.5.2.2 – Semantic μοντέλο για το debates dataset . **Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.5.3.1 – Περιβάλλον Apache Jena Fuseki..... **Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.5.3.2 – Δείγμα απάντησης στο Apache Jena Fuseki σε μορφή πίνακα**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 3.6.1 – Απόσπασμα από κύριο μέρος ενός αρχείου xml-tei ParlaMint**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 4.2.1 – Παράδειγμα χρήσης – ερώτημα 1 με «Αλέξιος Τσίπρα»**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 4.2.2 – Παράδειγμα χρήσης – αποτέλεσμα στο ερώτημα 1 με «Αλέξιος Τσίπρα»**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 4.2.3 – Παράδειγμα χρήσης – ερώτημα 2 με βουλευτές Νέας Δημοκρατίας**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Εικόνα 4.2.4 – Παράδειγμα χρήσης – αποτέλεσμα στο ερώτημα 2 με βουλευτές της Νέας Δημοκρατίας **Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**

Κατάλογος Πινάκων

Πίνακας 4.1.1 – Η εξέλιξη του Διαδικτύου 1

Πίνακας 2.2.1 – The Linked Open Data Cloud 1

Πίνακας 2.2.2 – Ανάπτυξη των συνδεδεμένων ανοικτών δεδομένων από το 2007 1

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 ιστοτόπου υποστηρίζει κυρίως την παραδοσιακή περιήγηση και αναζήτηση με βάση κυρίως 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 έλλειψη δομημένης αναπαράστασης δεδομένων. Οι συζητήσεις παρουσιάζονται κυρίως ως αδόμητο κείμενο, σε μορφή pdf, doc(x) ή txt, γεγονός που καθιστά δύσκολη την εξαγωγή

συγκεκριμένων πληροφοριών ή την εκτέλεση ουσιαστικής ανάλυσης σε λεπτομερές επίπεδο. Αυτός ο περιορισμός εμποδίζει την ολοκληρωμένη έρευνα, τον εντοπισμό τάσεων και την ικανότητα να γίνουν συνδέσεις μεταξύ διαφορετικών συζητήσεων, θεμάτων ή συμμετεχόντων/ομιλητών.

Η διπλωματική αυτή αποσκοπεί στην αντιμετώπιση αυτών των προκλήσεων προτείνοντας ένα πλαίσιο που μετατρέπει τις συζητήσεις του Ελληνικού Κοινοβουλίου σε συνδεδεμένα δεδομένα χρησιμοποιώντας τεχνολογίες σημασιολογικού ιστού. Με την αξιοποίηση της γλώσσας προγραμματισμού Python και του ANTLR-4, τα μη δομημένα κείμενα θα αναλυθούν και θα μετατραπούν σε δομημένα έγγραφα XML. Στη συνέχεια, τα έγγραφα XML θα μετατραπούν σε μορφή σχήματος RDFS (Resource Description Framework Schema), επιτρέποντας την αναπαράσταση εννοιών, σχέσεων και μεταδεδομένων που σχετίζονται με τις συζητήσεις.

Ο πρωταρχικός στόχος είναι να δημιουργηθεί μια αναπαράσταση συνδεδεμένων δεδομένων των συζητήσεων του Ελληνικού Κοινοβουλίου, η οποία θα επιτρέπει τη σημασιολογική αναζήτηση μέσω SPARQL. Αυτή η προσέγγιση θα διευκολύνει τις προηγμένες αναζητήσεις, τα σύνθετα ερωτήματα και την εξαγωγή ουσιαστικών πληροφοριών από το σύνολο δεδομένων. Επιπλέον, η εργασία θα διερευνήσει τη δημιουργία ενός σχήματος RDFS ειδικά για τον τομέα των κοινοβουλευτικών συζητήσεων της Ελλάδος, αποτυπώνοντας την απαραίτητη γνώση για ολοκληρωμένη ανάλυση.

Συνολικά, η παρούσα διπλωματική θα μπορούσαμε να πούμε πως επιδιώκει να συμβάλει στον τομέα της ηλεκτρονικής δημοκρατίας και της διαφανούς διακυβέρνησης αναδεικνύοντας τις δυνατότητες των τεχνολογιών του σημασιολογικού ιστού στις κοινοβουλευτικές συζητήσεις.

1.2 Αντικείμενο διπλωματικής

Εδώ αναφερόμαστε συγκεκριμένα στο τί θα κάνει η διπλωματική. Αναφέρουμε λεπτομερώς α) τα προβλήματα που θα λύσει (και που ήδη έχουν περιγραφεί γενικά στην προηγούμενη ενότητα), και β) πώς σκοπεύει να τα λύσει.

Είναι σημαντικό κάποιος που θα διαβάσει την ενότητα αυτή να καταλάβει σε σημαντικό βαθμό τον σκοπό της διπλωματικής σας και τις τεχνικές δυσκολίες της, χωρίς να είναι αναγκαίο να δει όλα τα άλλα κεφάλαια. Η ενότητα αυτή θέλει πολύ προσοχή και καλύτερα να τη γράψετε αφού έχετε γράψει όλα τα υπόλοιπα κεφάλαια.

1.2.1 Συνεισφορά

Εδώ παραθέτουμε αριθμητικά συγκεκριμένες ενέργειες/λύσεις/μεθοδολογίες που παρουσιάζει η διπλωματική και λύνουν τα προβλήματα που υποσχεθήκαμε στην προηγούμενη ενότητα ότι θα λύσει η διπλωματική. Συνήθως η υποενότητα αυτή έχει την παρακάτω μορφή:

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. **Μελετήσαμε συστήματα κ.λ.π.**
2. **Υλοποιήσαμε 3 αλγορίθμους υπολογισμού κ.λ.π.**
3. **Αξιολογήσαμε την επίδοση των αλγορίθμων και βρήκαμε ότι κ.λ.π.**
4. **Ενσωματώσαμε τους αλγορίθμους σε πρότυπο σύστημα κ.λ.π.**
5. **...**

1.3 Σχετικά Παραδείγματα του Εξωτερικού

Είναι ζωτικής σημασίας να εξετάσουμε σχετικά διεθνή παραδείγματα προκειμένου να κατανοήσουμε την εφαρμογή των συνδεδεμένων δεδομένων στις κοινοβουλευτικές διαδικασίες από παγκόσμια άποψη. 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 Κοινοβουλίου είναι ένα αξιοσημείωτο έργο που χρησιμοποιεί τις αρχές των συνδεδεμένων δεδομένων για να παρέχει πρόσβαση στις κοινοβουλευτικές συζητήσεις και τις σχετικές πληροφορίες (https://linkedpolitics.ops.few.vu.nl/web/html/home.html). Οι χρήστες μπορούν να περιηγηθούν και να εξετάσουν συζητήσεις, ομιλίες, αρχεία ψηφοφορίας και άλλες νομοθετικές δραστηριότητες στον ιστότοπο, ο οποίος 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 συνεδριάσει, 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 Καναδά (https://hansard.opennwt.ca/debates/) 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.

1.4 Οργάνωση κειμένου

Εδώ περιγράφουμε τα κεφάλαια της διπλωματικής: 1 πρόταση για το τι θα έχει κάθε κεφάλαιο. Συνήθως η ενότητα αυτή έχει την παρακάτω μορφή (δεν θα σας πάρει πάνω από 1 μεγάλη παράγραφο):

Εργασίες σχετικές με το αντικείμενο της διπλωματικής παρουσιάζονται στο Κεφάλαιο 2 . Το Κεφάλαιο 3 συζητά θέματα μοντελοποίησης. Στο Κεφάλαιο 4 αναπτύσσουμε κ.λ.π.

2

Θεωρητικό υπόβαθρο

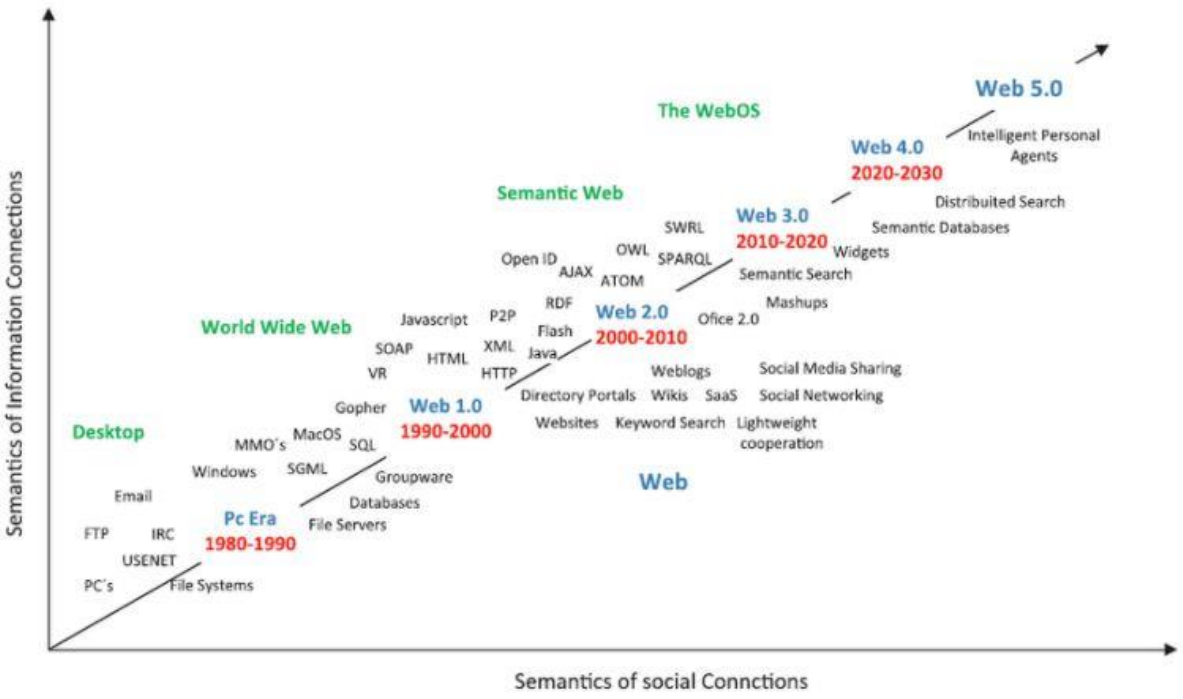
Σε αυτό το κεφάλαιο της εργασίας, θα γίνει μια εκτενής ανάλυση των βασικών θεωρητικών εννοιών που περικλείονται στο νοηματικό περιεχόμενο της εργασίας, και πιο συγκεκριμένα θα εστιάσουμε στον Σημασιολογικού Ιστού και στα βασικά συστατικά του. Αρχικά γίνεται μια ενδελεχή παρουσίαση του Σημασιολογικού Ιστού και στην συνέχεια εμβαθύνουμε στον τομέα των Ανοικτών Συνδεδεμένων Δεδομένων. Έπειτα, αναλύονται βασικές τεχνολογίες του Σημασιολογικού Ιστού, εστιάζοντας στην πολυπλοκότητα των RDF, XML, στην πολυμορφία των λεξικών και φυσικά στα ερωτήματα SPARQL. Τέλος, ολοκληρώνουμε με μια σύντομη αναφορά στα νομικών ανοικτών συνδεδεμένων δεδομένων, ένα πεδίο μελέτης στο οποίο οι αρχές του Σημασιολογικού Ιστού αλληλεπιδρούν με τις νομικές πληροφορίες.

2.1 Σημασιολογικός Ιστός – Semantic Web

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

Η μετάβαση από τον παραδοσιακό ιστό 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 2.1.1 Η εξέλιξη του Διαδικτύου (Osorio, Ortiz 2013 - https://myeltcafe.com/wp-content/uploads/2020/12/123123123.jpg)

Προφανώς, η μετάβαση στον σημασιολογικό ιστό υπόσχεται πολλά 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 ηλεκτρονικό εμπόριο, 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 ενοποίηση των ιατρικών δεδομένων, οδηγώντας σε βελτιωμένη περίθαλψη των ασθενών, ερευνητική συνεργασία και λήψη αποφάσεων βάσει δεδομένων.

Επιπλέον, ο σημασιολογικός ιστός έχει επιπτώσεις 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 ευφυών πρακτόρων και chatbots που μπορούν να κατανοούν και να απαντούν σε ερωτήματα χρηστών με πιο φυσικό και συνειδητό τρόπο.

Πλέον, μοναδικός στόχος της τεχνολογίας είναι η πρόοδος της, επομένως όπως μας μαρτυρά και η Εικόνα 2.1.1, είναι ότι πλησιάζει η εδραίωση του Semantic Web ή του Web 3.0 και η σταδιακή μας μετάβαση στο νεότερο διαδίκτυο με αρίθμηση 4.0, αν βέβαια επιβεβαιωθούν αυτές οι προβλέψεις, μέχρι το τέλος αυτής της δεκαετίας.

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 Ανοικτά 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 ανοιχτών 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.

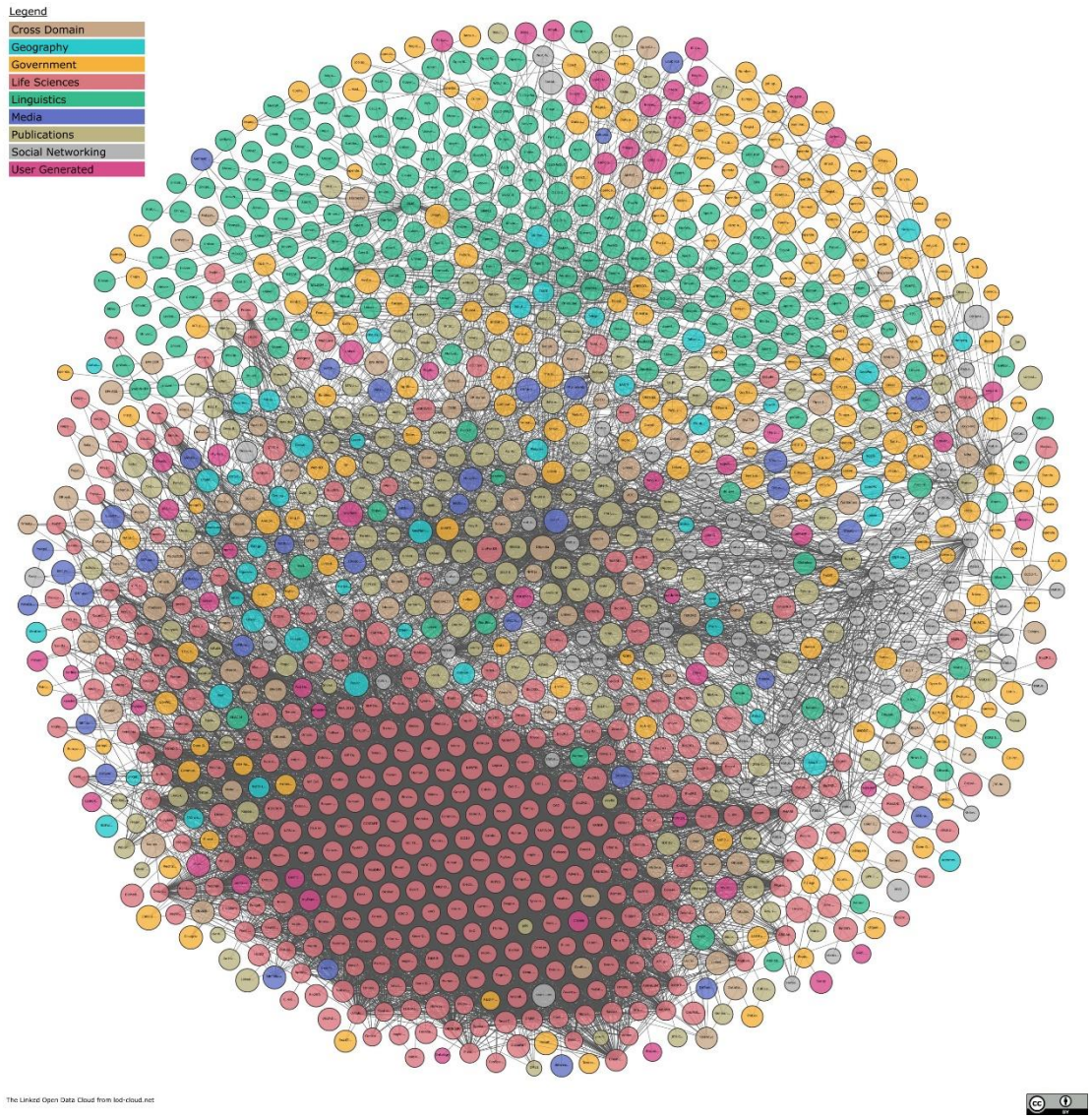
Η χρήση ομοιόμορφων αναγνωριστικών πόρων (URI) για τον ειδικό προσδιορισμό πραγμάτων και εννοιών εντός του οικο-συστήματος των συνδεδεμένων δεδομένων επιβάλλεται επίσης από τις αρχές των ανοικτών συνδεδεμένων δεδομένων. Τα URI χρησιμεύουν ως μόνιμα και παγκοσμίως μοναδικά αναγνωριστικά, επιτρέποντας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

Όλα όσα περιγράψαμε παραπάνω έχει συμπυκνώσει ο Τιμ Μπέρνερς Λι, ο εφευρέτης του Παγκόσμιου Ιστού, σε τέσσερις βασικές και θεμελιώδης αρχές. 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.

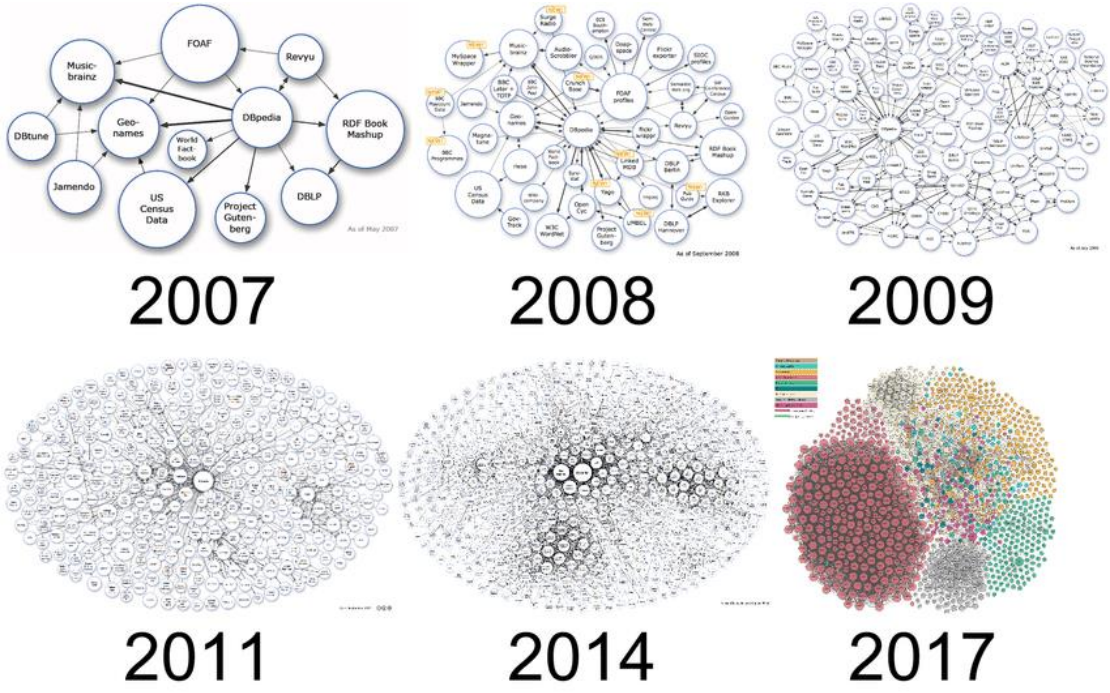
- Ο δεύτερος κανόνας καθιστά απλή την αναζήτηση αυτών των ονομάτων με τη χρήση HTTP URIs, επιτρέποντας στους χρήστες να έχουν πρόσβαση και να ανακτούν χρήσιμα δεδομένα που σχετίζονται με τους αναγνωρισμένους πόρους.
- Για την επιτυχή διευθέτηση και αναζήτηση των δεδομένων, ο τρίτος κανόνας δίνει έμφαση 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 προωθεί τη διασύνδεση και την ενσωμάτωση διαφορετικών συνόλων δεδομένων στο διαδίκτυο, όπως φαίνεται από την Εικόνα 2.2.1 (LOD Cloud). Στο διάγραμμα αυτό αποτυπώνεται το μεγαλειώδες δίκτυο των συνδεδεμένων συνόλων δεδομένων, τα οποία εμφανίζονται ως κόμβοι, και τις συνδέσεις 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.



Σχήμα 2.2.1 – The Linked Open Data Cloud (http://cas.lod-cloud.net/clouds/lod-cloud-sm.jpg)

Ακόμα ένα άλλο ενδιαφέρον γράφημα είναι αυτό της Εικόνας 2.2.2 στο οποίο 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 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 2.2.2 – Ανάπτυξη των συνδεδεμένων ανοικτών δεδομένων από το 2007 (https://www.researchgate.net/figure/Growth-of-Linked-Open-Data-since-2007-1-The-amount-of-data-sets-published-as-LOD-have_fig2_331748480)

2.3 Τεχνολογίες Υποστήριξης Δεδομένων

Στον τομέα του Σημασιολογικού Ιστού και των Συνδεδεμένων Δεδομένων περιλαμβάνονται διάφορες τεχνολογίες που είναι ζωτικής σημασίας για την αποτελεσματική αναπαράσταση, ολοκλήρωση, επερώτηση 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 (SPARQL Protocol and RDF Query Language), 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 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 αφετηρία 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 Ιστού (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>
```

Εικόνα 2.3.1 – Δείγμα Κώδικα XML

Στο παράδειγμα της εικόνας 2.3.1, 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 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.

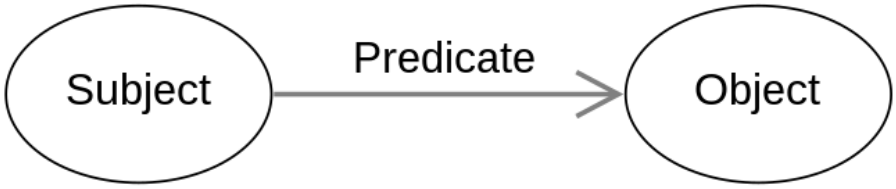
Στο πλαίσιο των συνδεδεμένων δεδομένων, η XML αποτελεί βασική τεχνική για τη δόμηση και τη δημιουργία εγγράφων αναγνώσιμων από μηχανήματα για μη δομημένα κείμενα. Αυτά τα έγγραφα μπορούν να επεξεργαστούν, να ταξινομηθούν και να επισημανθούν με χρήσιμα στοιχεία και χαρακτηριστικά με τη χρήση XML, με αποτέλεσμα την παραγωγή καλά διαμορφωμένων εγγράφων XML. Τα επόμενα βήματα της μετάφρασης, της ολοκλήρωσης 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, γνωστό ως RDF, είναι μια θεμελιώδης τεχνολογία στον Σημασιολογικό Ιστό που χρησιμεύει ως τυποποιημένο 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. Οι πόροι που περιγράφονται αντιπροσωπεύονται από το υποκείμενο, το κατ ηγόρημα 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.



Εικόνα 2.3.2.1 – Βασική σύνταξη RDF
(https://upload.wikimedia.org/wikipedia/commons/thumb/8/88/Basic_RDF_Graph.svg/800px-Basic_RDF_Graph.svg.png)

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.

Δεύτερον, το RDF διευκολύνει την έκφραση σχέσεων και σημασιολογίας μέσα στα δεδομένα. Το RDF παρέχει παγκοσμίως μοναδικά και μόνιμα αναγνωριστικά για τους πόρους χρησιμοποιώντας τα URI ως αναγνωριστικά. 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.

Χρησιμοποιώντας οντολογίες, το RDF διευκολύνει επίσης την επεκτασιμότητα. Οι έννοιες, οι συνδέσεις και οι περιορισμοί σε έναν τομέα αναπαρίστανται τυπικά και δομικά με οντολογίες. Η γνώση που αφορά συγκεκριμένο τομέα μπορεί να κωδικοποιηθεί χρησιμοποιώντας γλώσσες όπως το RDF Schema (RDFS) ή η Web Ontology Language (OWL).

Τέλος, τα δεδομένα RDF μπορούν να αποθηκευτούν και να αναζητηθούν αποτελεσματικά με τη χρήση του Apache Fuseki, μιας βάσης δεδομένων RDF ανοικτού κώδικα και ενός SPARQL endpoint. 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 μόνιμης αποθήκευσης δεδομένων και της βελτιστοποίησης ερωτημάτων.

Επιπλέον, το Fuseki προσφέρει ένα 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. Να τονίσουμε πως η λειτουργία και η σύνταξη των ερωτημάτων SPARQL θα αναλυθεί λεπτομερώς σε επόμενη παράγραφο.

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 έναν συγκεκριμένο τομέα. 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 προσφέρει ένα σύνολο κλάσεων και ιδιοτήτων για τη δημιουργία βασικών οντολογιών. Το RDFS παρέχει τα θεμέλια για τον ορισμό κλάσεων, ιδιοτήτων και σχέσεων μεταξύ τους, επιτρέποντας την ιεραρχική ταξινόμηση, τον προσδιορισμό ιδιοτήτων και τους περιορισμούς τομέα/περιοχής.

Ένα άλλο ισχυρό λεξιλόγιο είναι το Web Ontology Language (OWL), το οποίο παρέχει έναν πιο επίσημο και εκφραστικό τρόπο αναπαράστασης της γνώσης, επιτρέποντας τη διατύπωση περίπλοκων ιδεών, αξιωμάτων και λογικών επιχειρημάτων. Καθιστά δυνατή τη δημιουργία περίπλοκων οντολογιών που υποστηρίζουν πολύπλοκους συλλογισμούς και

αυτοματοποιημένη εξαγωγή συμπερασμάτων, γεγονός που προάγει τις δυνατότητες αναπαράστασης γνώσης και εξαγωγής συμπερασμάτων.

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

- FOAF (Friend of a Friend):

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

- SKOS (Simple Knowledge Organization System)

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

- Dublin Core

  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.

- Schema.org

  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.

- DBpedia

  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.

2.3.4 SPARQL

Η SPARQL (SPARQL Protocol and RDF Query Language) είναι μια ισχυρή 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 δεδομένων που είναι αποθηκευμένα ως RDF δεδομένα.

Τα ερωτήματα των 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 θα επιστραφούν 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 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.

Η SPARQL διαδραματίζει πρωταγωνιστικό ρόλο στο Apache 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 – Akoma Ntoso

Τα διασυνδεδεμένα ανοικτά κυβερνητικά δεδομένα, σε συντομογραφία 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 περιλαμβάνει 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, το οποίο χρησιμοποιεί τα μοναδικά αναγνωριστικά (URIs). Τα κυβερνητικά

δεδομένα ορίζονται ως διασυνδεδεμένα όταν ακολουθούνται αυτές τις αρχές, καθιστώντας παράλληλα απλή τη σύνδεση και την ανάμειξή τους με άλλα σύνολα δεδομένων για τη δημιουργία ενός πλούσιου ιστού διασυνδεδεμένης γνώσης.

Η υιοθέτηση του 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, το LOGD ενθαρρύνει 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. Το LOGD διευκολύνει 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 κοινωνικών προβλημάτων.

Κυβερνήσεις σε όλο τον κόσμο έχουν αναλάβει την «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 χάρη στο Akoma Ntoso, επίσης γνωστό ως LegalDocML, το οποίο είναι ένα πρότυπο βασισμένο στην XML. Το Akoma Ntoso δημιουργεί έναν δομημένο μορφότυπο για την απεικόνιση νομικών εγγράφων που υποστηρίζει τη διαλειτουργικότητα, την προσβασιμότητα και τη σημασιολογία, βασιζόμενο στις ιδέες των Συνδεδεμένων Ανοικτών Δεδομένων.

Πιο συγκεκριμένα, το Akoma Ntoso παρουσιάζει μια ενδελεχή συλλογή στοιχείων και χαρακτηριστικών που είναι ειδικά σχεδιασμένα για την αποτύπωση της μορφής και του περιεχομένου των νομοθετικών 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 Akoma Ntoso, απλοποιώντας την ανταλλαγή, την ανάλυση και την επαναχρησιμοποίησή τους.

Πρόσθετα οφέλη προκύπτουν από την ενσωμάτωση του Akoma Ntoso με τα Συνδεδεμένα Ανοικτά Κυβερνητικά Δεδομένα. Το Akoma Ntoso, ως αυστηρή και τυποποιημένη μορφή παρουσίασης των δεδομένων, καθιστά δυνατή τη διασύνδεση νομικών εγγράφων με άλλα συναφή δεδομένα και σύνολα δεδομένων, αξιοποιώντας τις έννοιες των Συνδεδεμένων Δεδομένων. Με τη χρήση αυτής της δυνατότητας, τα νομικά κείμενα μπορούν να ενσωματωθούν απρόσκοπτα με μεταδεδομένα, δικαστικές αποφάσεις, κανόνες και άλλες σχετικές πηγές δεδομένων, δημιουργώντας ένα πιο λεπτομερές και ολοκληρωμένο γράφημα νομικής γνώσης.

Το Akoma Ntoso χρησιμοποιείται τόσο για την βελτίωση της προσβασιμότητας των νομικών 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 νομικών κειμένων στο Akoma Ntoso. Εκτός από τη δυνατότητα ταχύτερης πρόσβασης σε σχετικές πληροφορίες αυτό ενθαρρύνει κυρίως τη διαφάνεια, την αποτελεσματικότητα και την ομοιομορφία των νομικών διαδικασιών.

Κυβερνήσεις, δικαστήρια και άλλοι οργανισμοί σε όλο τον κόσμο έχουν υιοθετήσει ευρέως το Akoma Ntoso ως πρότυπο για την παρουσίαση νομοθετικών κειμένων. Η χρήση του διασφαλίζει ότι οι νομικές πληροφορίες είναι τυποποιημένες και διαλειτουργικές, προωθώντας τη διεθνή συνεργασία, τη συγκριτική νομική έρευνα και τη δημιουργία νομικών οντολογιών και βάσεων γνώσης.

2.5 TEI ParlaMint

Όπως έχει γίνει αντιληπτό, η τυποποίηση μορφώτυπων δεδομένων και η διαλειτουργικότητα τους είναι ζωτικής σημασίας. Για τον σκοπό αυτό η ομάδα Text Encoding Initiative (TEI), σε συνεργασία με την CLARIN, ανάπτυξε ένα πρότυπο με βάση την XML για την έκφραση δεδομένων στον τομέα του κοινοβουλευτικού κλάδου, με χαρακτηριστική ονομασία «TEI ParlaMint». Το TEI ParlaMint προσφέρει ένα ισχυρό θεμέλιο για εμβάθυνση στην πολυπλοκότητα των νομοθετικών διαδικασιών, ενσωματώνοντας απρόσκοπτα δομημένα κοινοβουλευτικά δεδομένα.

Το XML Akoma Ntoso αναπαριστά νομικά έγγραφα με λεπτομερή ιεραρχική δομή. Σε αντίθεση, το TEI ParlaMint, μια επέκταση του προτύπου TEI, 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 το μορφότυπο TEI ώστε να προσαρμόζεται στις ιδιαιτερότητες των νομοθετικών δεδομένων, δημιουργώντας μια ενιαία πλατφόρμα για μελέτη και σύγκριση.

Πέρα από την εμφανή διαλειτουργικότητα, το TEI 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.

3

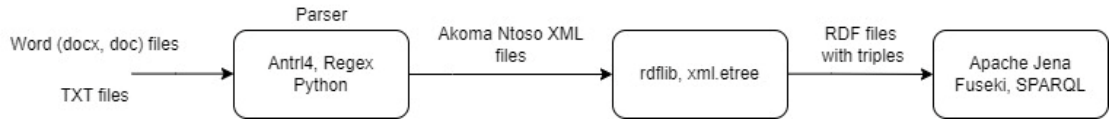
Σύστημα Διαχείρισης Πρακτικών

Η θεωρητική θεμελίωση που δημιουργήθηκε στις προηγούμενες ενότητες παρέχει το απαραίτητο υπόβαθρο για την πρακτική εφαρμογή της διπλωματικής εργασίας. Με βάση τις έννοιες του Σημασιολογικού Ιστού, των Συνδεδεμένων Δεδομένων και των συναφών τεχνολογιών, το πρακτικό μέρος θα επικεντρωθεί στην εφαρμογή αυτών των αρχών για τη μετατροπή των συζητήσεων του ελληνικού κοινοβουλίου σε ανοιχτά διασυνδεδεμένα δεδομένα.

Πιο συγκεκριμένα, θα διευρύνει τις πρακτικές πτυχές της διαδικασίας υλοποίησης χρησιμοποιώντας τη θεωρητική κατανόηση της XML, του RDF και του Akoma Ntoso. Θα περιγράφει ο τρόπο χρήσης του ANTLR4 και της γλώσσας Python για τη μετατροπή εγγράφων Word και κειμένου TXT σε αρχεία XML – Akoma Ntoso. Στη συνέχεια, αυτά τα αρχεία XML θα μετατραπούν σε τριπλέτες RDF. Η βάση δεδομένων Apache Fuseki RDF, η οποία θα χρησιμοποιηθεί ως backend για τη φιλοξενία και την αναζήτηση των συνδεδεμένων δεδομένων, θα γεμίσει με τα δεδομένα RDF. Τέλος, θα εξεταστεί και ο τρόπος χρήσης των αναζητήσεων SPARQL για την εξαγωγή ορισμένων δεδομένων από τα συνδεδεμένα δεδομένα.

3.1 Αρχιτεκτονική Συστήματος

Η αρχιτεκτονική του έργου περιλαμβάνει όλα τα στοιχεία και τις τεχνολογίες που απαιτούνται για τη μετατροπή των συζητήσεων του ελληνικού κοινοβουλίου σε συνδεδεμένα δεδομένα. Η αρχιτεκτονική χρησιμοποιεί μια μεθοδική διαδικασία για τον συνδυασμό διαφόρων σταδίων και εργαλείων, προκειμένου να παραχθούν τα επιθυμητά αποτελέσματα. Ακολουθεί μια πρώτη επαφή με τα στοιχεία της αρχιτεκτονικής του συστήματος, όπως αυτά αποτυπώνονται και στην εικόνα 3.1.1.



Εικόνα 3.1.1 – Διάγραμμα Αρχιτεκτονικής Συστήματος

Το πρώτο στάδιο 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 υπάρχον εργασίες.

Σε δεύτερό βήμα, πιο χρονοβόρο και πιο περίπλοκο από όλα, είναι η επεξεργασία αυτών των αρχικών ‘κινητήριων’ αρχείων. Αυτά θα πρέπει να μετατραπούν σε δομημένη μορφή XML, μέσω της αξιοποίησης ενός Parser, που βασίζεται στην ANTLR4 και στην Python γλώσσα. Ο parser αναλύει τα έγγραφα, και προσδιορίζοντας τα σχετικά στοιχεία και χαρακτηριστικά, παράγει αρχεία XML που αντιπροσωπεύουν τη δομημένη αναπαράσταση των συζητήσεων.

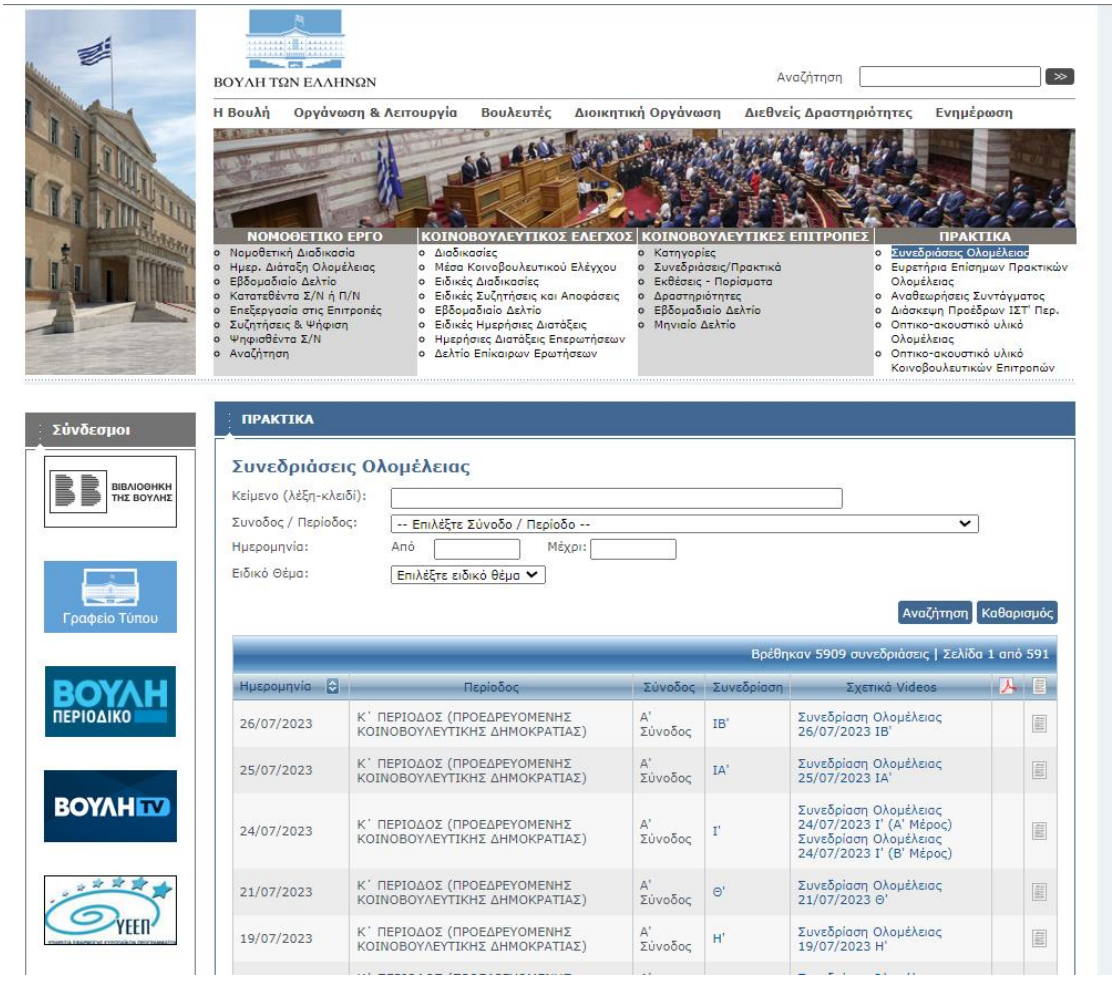
Στην συνέχεια, τα παραγόμενα αρχεία XML υποβάλλονται σε επεξεργασία και με κατάλληλο συνδυασμό βιβλιοθηκών μετασχηματίζονται σε τριπλέτες RDF. Ο μετασχηματισμός αυτός περιλαμβάνει την αντιστοίχιση των στοιχείων και χαρακτηριστικών XML σε ιδιότητες και πόρους RDF. Τα μετασχηματισμένα δεδομένα τηρούν τη μορφή RDF Schema (RDFS), εξασφαλίζοντας τη διαλειτουργικότητα και τη συμβατότητα με τις αρχές των συνδεδεμένων δεδομένων.

Τέλος, τα σύνολο των τριπλετών αποθηκεύονται και διαχειρίζονται σε μια βάση δεδομένων RDF. Στην δική μας περίπτωση επιλέχθηκε το Apache Jena Fuseki, μια δημοφιλής βάση δεδομένων RDF, που θα χρησιμεύσει για τη φιλοξενία των συνδεδεμένων δεδομένων. Το 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.

3.2 Βάση Δεδομένων

Το "Α" και το "Ω" ενός έργου, όπως αυτό που περιγράφεται στην παρόν εργασία, είναι μια λεπτομερής και ολοκληρωμένη βάση δεδομένων. Μια βάση δεδομένων που περιέχει αρχεία έγκυρα και πολυπληθή, δίνουν στο όλο σύστημα ένα πιο αξιόπιστο κύρος.

Η βάση δεδομένων μας περιέχει αρχεία που ανακτήθηκαν απευθείας από τον επίσημο ιστότοπο της Ελληνικής Κυβέρνησης (https://www.hellenicparliament.gr/Praktika/Synedriaseis-Olomeleias), καθιστώντας την αυτομάτως αξιόπιστη για εμπεριστατωμένη ανάλυση και έρευνα. Το γεγονός ότι το σύνολο των δεδομένων προέρχεται από μια επίσημη ιστοσελίδα ενισχύει τη νομιμότητα του έργου και επιτρέπει την έγκαιρη κατανόηση του συνεχώς μεταβαλλόμενου πολιτικού περιβάλλοντος της Ελλάδας. Μιας και το πεδίο μελέτης έχει οριοθετηθεί από τις αρχές της δεκαετίας του 1990 μέχρι και σήμερα, προκύπτει πως το σύνολο το αρχείων ανέρχεται περίπου στα 5900 αρχεία, μορφής word και txt.



Εικόνα 3.2.1 – Επίσημος ιστότοπος Ελληνικής Κυβέρνησης με Συνεδριάσεις Ολομέλειας

Επιπλέον, συμπληρωματικά αρχεία έχουν αντληθεί από ένα δημόσιο repository αφιερωμένο στις ελληνικές νομοθετικές διαδικασίες. Επιλέχθηκαν λοιπόν αρχεία *csv* που περιέχουν πληροφορίες για πολιτικές θητείες και περιόδους κυβερνήσεων, γεγονός που αποσκοπεί στο να αυξήσει περεταίρω το εύρος και το βάθος του συνόλου δεδομένων. Σε

γενικές γραμμές, το repository αυτό τιτλοφορείται ως "*Greek_Parliament_Proceedings_Dataset*" (https://github.com/Dritsa-Konstantina/Greek_Parliament_Proceedings_Dataset) και είναι ένα project που επεξεργάζεται τα αρχεία των συνεδριάσεων του Ελληνικού Κοινοβουλίου, σε μια άλλη διάσταση από αυτή τις παρούσας εργασίας, μιας και δεν εστιάζει καθόλου στην διασύνδεση των δεδομένων, αλλά επικεντρώνεται στην αξιολόγηση της ποιότητα του λόγου και στην εξέταση των συναισθημάτων που προκύπτουν από τα λεγόμενα των ομιλητών. Είναι ένα επιτυχημένο εγχείρημα, το όποιο έφερε εις πέρας η κ. *Δρίτσα Κωνσταντίνα* στο πλαίσιο της διδακτορικής εργασίας της στο Οικονομικό Πανεπιστήμιο Αθηνών, σε συνεργασία με τον μη κερδοσκοπικός δημοσιογραφικός οργανισμός *iMEdD*.

3.3 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 (αρχές δεκαετίας του 90 μέχρι και σήμερα) παρατηρείται μια εναλλαγή στον τρόπο συγγραφής τους. Για αυτό το λόγο η ανάλυση δεν μπορεί να είναι υπερπλήρης, αλλά καλύπτεται ένα μεγάλο μέρος των αρχείων, μιας και θα γίνει μια εκτεταμένη ανάλυση με αρκετές από τις υπο-κατηγορίες που εντοπίστηκαν.

Στην αρχή, συναντά κανείς τον «ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ» (Εικόνα 3.3.1) που περιλαμβάνει συνήθως βασικές πληροφορίες, όπως η περίοδος που καλύπτεται (π.χ. ΙΖ' ΠΕΡΙΟΔΟΣ), ο τύπος της κυβέρνησης (π.χ. ΠΡΟΕΔΡΙΚΗ ΚΟΙΝΟΒΟΥΛΕΥΤΙΚΗ ΔΗΜΟΚΡΑΤΙΑ), η σύνοδος (π.χ. Σύνοδος Γ') και ο αριθμός της συνεδρίασης (π.χ. ΣΥΝΕΔΡΙΑΣΗ ΡΓΛ'). Επιπλέον, περιέχει τη συγκεκριμένη ημερομηνία κατά την οποία έλαβε χώρα η κοινοβουλευτική σύνοδος, όπως "Δευτέρα, 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 3.3.1 – 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 3.3.2). 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

Α. 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 Γαλιάς 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 γραφείου του Βουλευτή κ. Μ. Βαρβιτσιώτη και καταδίκη αυτής, σελ.
6. Επί διαδικαστικού θέματος, σελ.

Β. 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 3.3.2 – Ενδεικτικό απόσπασμα πρακτικών – «Θεμάτων» (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 3.3.3). 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 γραφείου 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 Ά. , σελ.

Εικόνα 3.3.3 – 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» (Εικόνα 3.3.4). Αναφέρει εκ νέου την περίοδο, τον τύπο της κυβέρνησης, τη σύνοδο, τον αριθμό της συνεδρίασης και την ημερομηνία για λόγους αναφοράς.

ΠΡΑΚΤΙΚΑ ΒΟΥΛΗΣ

ΙΖ΄ ΠΕΡΙΟΔΟΣ

ΠΡΟΕΔΡΕΥΟΜΕΝΗΣ ΚΟΙΝΟΒΟΥΛΕΥΤΙΚΗΣ ΔΗΜΟΚΡΑΤΙΑΣ

ΣΥΝΟΔΟΣ Γ΄

ΣΥΝΕΔΡΙΑΣΗ ΡΛΓ΄

Παρασκευή 8 Ιουνίου 2018

Εικόνα 3.3.4 – Ενδεικτικό απόσπασμα πρακτικών – «Πρακτικά Βουλής» (08-06-2018)

Στη συνέχεια, υπάρχει ένας σύντομος πρόλογος (Εικόνα 3.3.5), 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.

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**.

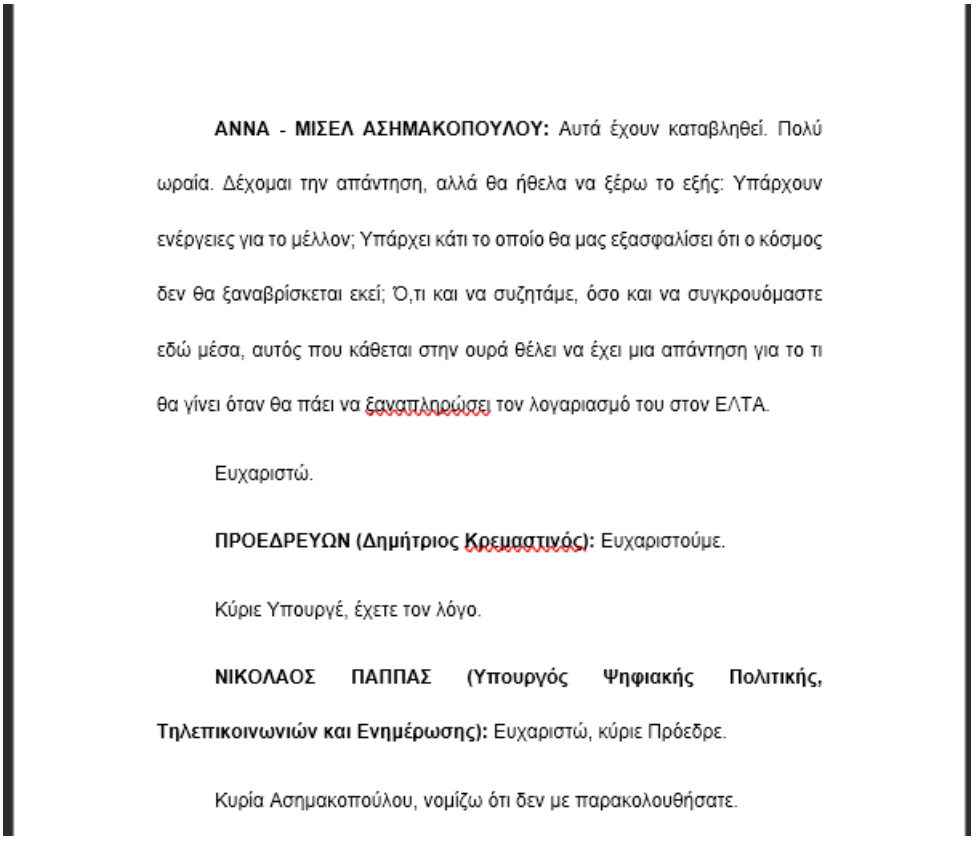
Εικόνα 3.3.5 – Ενδεικτικό απόσπασμα πρακτικών – Εισαγωγικός Πρόλογος (08-06-2018)

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 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 ενδιάμεσο και κύριο πυλώνα των πρακτικών, όπου είναι αυτό που περιέχονται οι διάλογοι χρησιμοποιείται 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 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 στην εικόνα 3.3.6 , 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. Σε αυτό το σημείο να σημειώσουμε πως η ανάλυση σε επόμενες ενότητες θα βασιστεί σε αυτό το κειμενικό απόσπασμα (Εικόνα 3.3.6) .



ΑΝΝΑ - 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 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.

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

3.4 Εξαγωγή και Ανάλυση Κειμένων

Η ανάλυση και η τροποποίηση των πρακτικών του Ελληνικού Κοινοβουλίου εξαρτάται σε μεγάλο βαθμό από το αρχείο που δημιουργήθηκε, με όνομα *convert_to_xml.py*, το οποίο λειτουργεί ως ένας parser. Ο parser αυτός επεξεργάζεται αποτελεσματικά τα ακατέργαστα αρχεία που βρίσκονται στον αρχικό φάκελο κάνοντας χρήση της γλώσσας ANTLR4 και των κανονικών εκφράσεων (regular expressions), τα οποία αναγνωρίζουν με ακρίβεια και εξάγουν τμήματα του αρχείου.

Τα εισαγωγικά τμήματα διακρίνονται από το κύριο σώμα των πρακτικών στην αρχή της διαδικασίας ανάλυσης από την ANTLR4. Η σύνοδος, η ημερομηνία και ο αριθμός της συνεδρίασης είναι μεταξύ των κρίσιμων μεταδεδομένων που αναγνωρίζει και ανακτά σε πρώτο στάδιο ο parser.

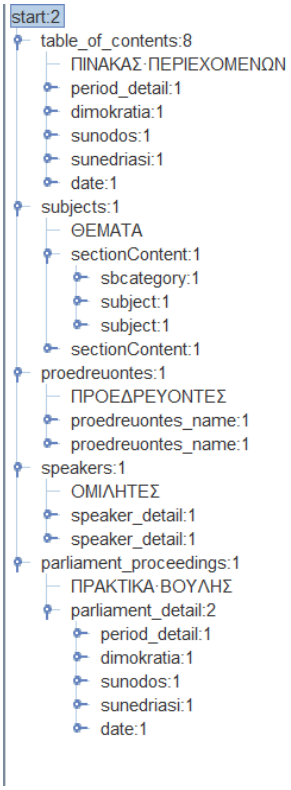
Ο parser μέσω γραμμών εντολών που αξιοποιούν «Κανονικές Εκφράσεις (Regular expression» ομαδοποιεί με ακρίβεια κάθε ομιλητή με τον αντίστοιχο λόγο του. Ο αναλυτής προσδιορίζει με αξιοπιστία το όνομα του ομιλητή και οριοθετεί τα όρια των ομιλιών του μέσα στις κοινοβουλετικές διαδικασίες με τη χρήση μοτίβων regex. ~~Κάθε λόγος αντιμετωπίζεται ως διακριτή μονάδα λόγω της αυστηρής τμηματοποίησης, η οποία διευκολύνει την πρόσθετη επεξεργασία και τον μετασχηματισμό~~. Η αξιοποίηση των regex εκφράσεων επικεντρώνεται κυρίως στην αναγνώριση του ομιλητή, του προέδρου ή άλλων σχόλιων όπως χειροκροτημάτων, ηλεκτρονικής καταμέτρησης. Σε όποιο σημείο χρησιμοποιούνται για την αναγνώριση κανονικές εκφράσεις, είναι λέξεις-φράσεις όπου ακολουθούν μια συγκεκριμένη δομή ή μορφοποίηση στην έκφραση τους, επομένως η τυποποίηση τους είναι δεδομένη, μιας και επιτυγχάνεται η αντιστοίχιση συγκεκριμενων μοτίβων.

Η επιτυχής ενσωμάτωση του αρχείου Python με το ANTLR4 και το Regex αναδεικνύει τον κρίσιμο ρόλο του στην επιτάχυνση της διαδικασίας μετατροπής και την αυτοματοποίηση της μετατροπής των νομοθετικών αρχείων σε συνδεδεμένα δεδομένα. Ο αναλυτής είναι ένα ισχυρό εργαλείο που διαχειρίζεται αποτελεσματικά την πολυπλοκότητα των διαδικασιών, διατηρώντας την ακρίβεια των εξαγόμενων δεδομένων και συμβάλλοντας στην παροχή μιας συνδεδεμένης, συνεκτικής απεικόνισης των συζητήσεων στο ελληνικό κοινοβούλιο.

3.4.1 ANLTR4 – REGEX

Η ANTLR (ANother Tool for Language Recognition), και πιο συγκεκριμένα η τέταρτη έκδοση, είναι μια ισχυρή γλώσσα η οποία χρησιμοποιείται κυρίως για τη δημιουργία αναλυτών για διάφορες γλώσσες προγραμματισμού, γλώσσες ειδικού τομέα (DSLs), ακόμη και μορφές δεδομένων. Είναι γνωστό για την ευκολία χρήσης, την αποδοτικότητα και την επεκτασιμότητά του, καθιστώντας το μια προτιμώμενη επιλογή σε έργα που απαιτούν δυνατότητες γλωσσικής επεξεργασίας.

Τα βασικά αρχεία τα όποια πρέπει να δημιουργηθούν για την ολοκληρωμένη λειτουργεία της γλώσσας είναι μία πλήρη γραμματική, 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 αυτό του σχήματος (Σχήμα 3.4.1.1), το οποίο περιέχει αρκετές από τις υποκατηγορίες που δύνανται να αναγνωριστούν. Φαίνεται πως κάθε στοιχείο έχει αναγνωριστεί από τον parser και έχει αντιστοιχηθεί στην συγκεκριμένη ετικέτα-μεταβλητή.



Εικόνα 3.4.1.1 – Ενδεικτικό δέντρο που προκύπτει από την γραμματική ANLTR4

```
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'
            | 'ΠΡΕΟΔΡΕΥΟΝΤΕΣ'
            | 'ΠΡΟΕΔΡΕΥΩΝ'
            | 'ΠΡΟΕΔΡΕΥΟΥΣΑ'
            | 'ΠΡΟΕΔΡΟΥΣΑ'
            | 'ΠΡΟΕΔΡΟΣ'
        ) SPACES
    );
```

Εικόνα 3.4.1.2 – Απόσπασμα από το συντακτικό της γραμματική

Στην παραπάνω φωτογραφία, στην εικόνα 3.4.1.2, παρατίθεται ένα μέρος του συντακτικού της γραμματικής μας. Φαίνεται πως η σύνταξη που ακολουθείται είναι ιδιαίτερη, και ονομάζεται *Backus–Naur form.* Ενδεικτικά, φαίνεται οι κανόνες για την σημείωση, τον πίνακα περιεχόμενων, για τα θέματα, για τους ομιλητές και για τους προεδρεύοντες.

3.5 Διαχείριση Δεδομένων

Ακολουθεί η μελέτη του βασικού ‘ταξιδιού’ για την μετατροπή των αρχείων σε ανοιχτά διασυνδεδεμένα δεδομένα και πιο συγκεκριμένα σε αρχεία RDF.

3.5.1 Μετατροπή Αρχείων Κειμένου σε XML αρχεία

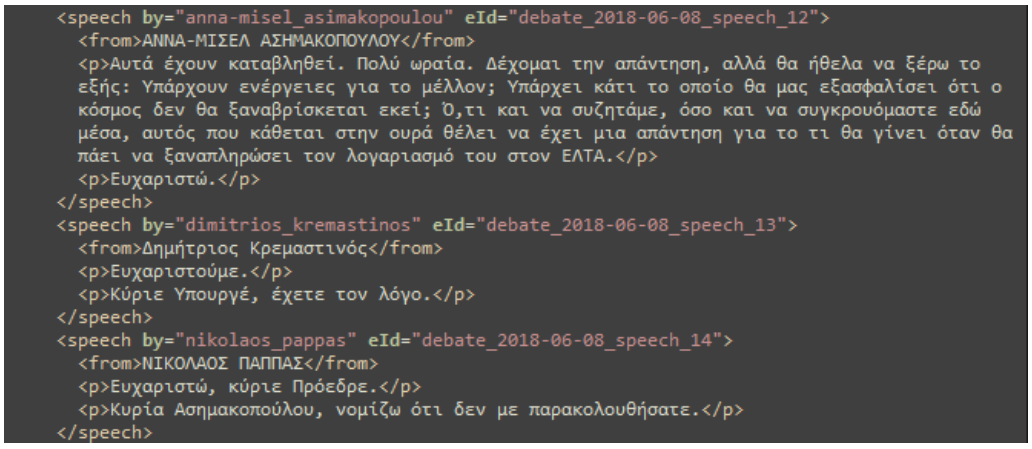
Σε πρώτο βήμα, καλούμαστε να δημιουργήσουμε αρχεία xml. Επιλέχθηκε αυτά τα αρχεία να ακολουθούν τις απαιτήσεις του Akoma Ntoso. Το αρχείο σε γλώσσα Python, που είναι υπεύθυνο για την πρώτη επεξεργασία αρχείων, όπως αναλύθηκε παραπάνω, εξάγει βασικά δεδομένα από τις κοινοβουλευτικές συνεδριάσεις, συμπεριλαμβανομένων των εισαγωγικών πληροφοριών, των ονομάτων των ομιλητών και του περιεχομένου των ομιλιών τους. Αυτές οι μεταβλητές που εξάγονται αποτελούν τη βάση για τη δημιουργία οργανωμένων αρχείων XML, που συμμορφώνονται με τα πρότυπα του Akoma Ntoso.

Τα παραγόμενα αρχεία XML εξασφαλίζουν μια ομοιόμορφη αναπαράσταση των νομοθετικών συζητήσεων με τη μεθοδική δόμηση XML σύμφωνα με τις κατευθυντήριες γραμμές Akoma Ntoso, όπως χαρακτηριστικά φαίνονται στην εικόνα 3.5.1.3 όπου δείχνει την ιεραρχική δομή που θα πρέπει να έχει κάθε αρχείο – debate. Όλα τα δεδομένα των αρχικών αρχείων αντιστοιχίζονται στα κατάλληλα στοιχεία XML.

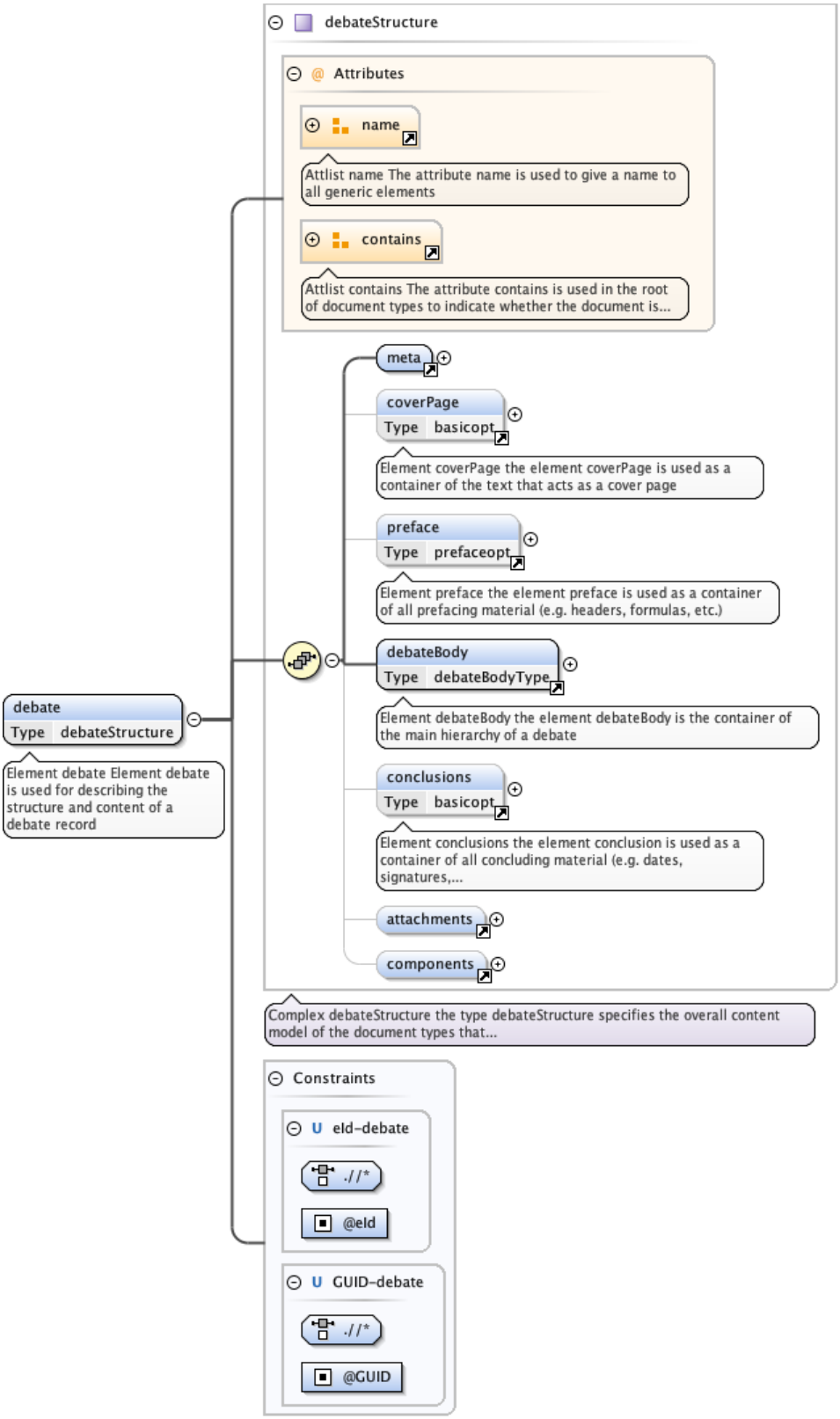
Πιο συγκεκριμένα, το μέρος των εισαγωγικών πληροφοριών, όπως ημερομηνία αριθμός συνόδου κ.α. χρησιμοποιούνται σαν μεταδεδομενα για τα αρχεία xml και τοποθετούνται κατάλληλα εντός της ετικέτας “<*meta*>”. Ενδεικτικά, να αναφέρουμε πως μέρος των μεταδεδομένων αποτελεί η λίστα με όλους του ομιλητές της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

Εικόνα 3.5.1.1 – Απόσπασμα από τα μεταδεδομένα ενός αρχείου xml-akoma ntoso.

Στην συνέχεια, φαίνεται το τελικό αποτέλεσμα ενός αρχείου xml για το κύριος μέρος του πρακτικού μιας συνεδρίασης. Αποτελεί συνέχεια της “εικόνας 3.3.6” στην φαίνονται οι ίδιοι διάλογοι ανάμεσα στους τρεις ομιλητές (Εικόνα 3.5.1.2).



Εικόνα 3.5.1.2 – Απόσπασμα από κύριο μέρος (debateBody) ενός αρχείου xml-akoma ntoso.



Εικόνα 3.5.1.3 – Διάγραμμα για το στοιχείο “debate” (http://docs.oasis-open.org/legaldocml/akn-core/v1.0/os/part2-specs/img/akomantoso30_xsd_Element_debate.png)

3.5.2 Μετατροπή Αρχείων XML σε RDF αρχεία

Σε αυτό το βήμα της εργασίας, βρίσκεται η διαδικασία μετατροπής των αρχείων xml σε μορφής RDF. Αυτό διευκολύνεται μέσω της χρήσης του 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 αρχεία Akoma Ntoso XML που παράχθηκαν στο προηγούμενο στάδιο, όσο και τα αρχεία csv από το repository της εργασίας της διδακτορικής φοιτήτριας του ΟΠΑ.

Πιο αναλυτικά, τα αρχεία XML αποτελούνται από στοιχεία με ετικέτες που αντικατοπτρίζουν διάφορα μέρη του αρχείου, 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 μετατρέπονται 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 ζωτικής σημασίας για την κατανόηση του πολιτικού τοπίου και της κατανομής των αρμοδιοτήτων μεταξύ των μελών του κοινοβουλίου.

Στα πλαίσια αυτής εργασίας επιλέχθηκε να μετατραπούν τα αρχεία rdf σε μορφή rdf/xml, όπως φαίνεται και στις εικόνες 3.5.2.1 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 έναν λόγο *(debate_2018-06-85_speech_12)*, στο (β) για έναν ομιλητή (GRmember_996), στο (γ) για πολιτικές θητείες και αξιώματα για τον ίδιο ομιλητή και τέλος στο (δ) για μία πολιτική θητεία *(political_tenure_359)*.

```
<rdf:Description rdf:about="https://purl.org/greekparldebates/debate_2018-06-08_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/speech_2018-06-08_debate_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 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.
Ευχαριστώ.</greek_lp:spokenText>
  </rdf:Description>
```

Εικόνα 3.5.2.1 (α)

```
</rdf:Description>
<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 3.5.2.1 (β)

```
<rdf:Description rdf:about="https://purl.org/greekparldebates/GRmember_996">
  <greek_lp:PoliticalTenure rdf:resource="https://purl.org/greekparldebates/political_tenure_359"/>
  <greek_lp:PoliticalTenure rdf:resource="https://purl.org/greekparldebates/political_tenure_360"/>
  <greek_lp:PoliticalTenure rdf:resource="https://purl.org/greekparldebates/political_tenure_361"/>
</rdf:Description>
```

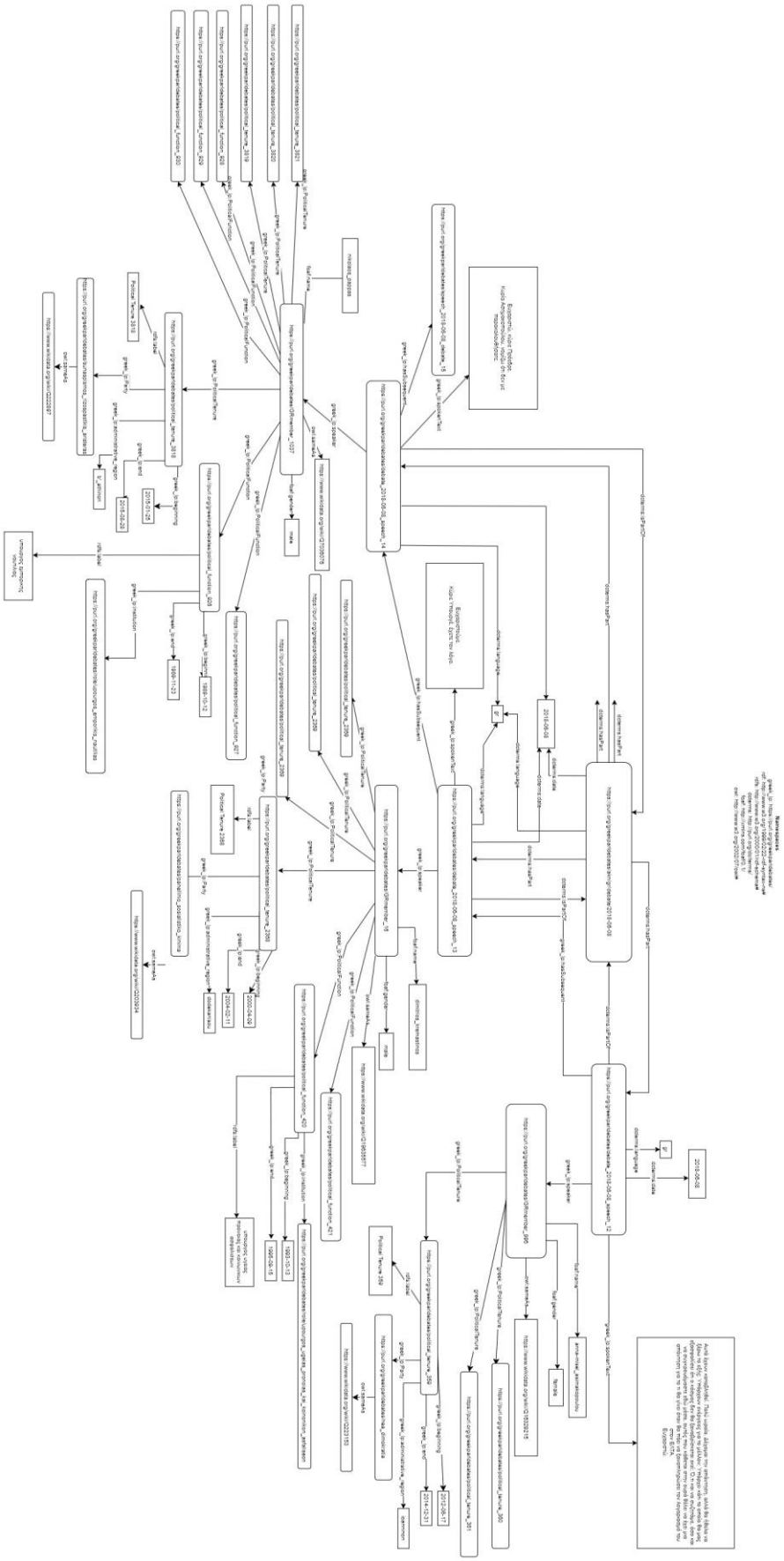
Εικόνα 3.5.2.1 (γ)

```
<rdf:Description rdf:about="https://purl.org/greekparldebates/political_tenure_359">
  <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 359</rdfs:label>
</rdf:Description>
```

Εικόνα 3.5.2.1 (δ)

Εικόνα 3.5.2.1 – Μερικές τριπλέτες RDF όπως φαίνονται στα rdf/xml αρχεία.

Τέλος, μέσω της επόμενη εικόνα (εικόνα 3.5.2.2) μπορούμε να εμβαθύνουμε στο σύνολο δεδομένων των πρακτικών του ελληνικού κοινοβουλίου μέσω μιας εξελιγμένης αναπαράστασης του σημασιολογικού μοντέλου. Με την πρώτη ματιά, αυτή η εικόνα μπορεί να φαίνεται σαν ένα πολύπλοκο δίκτυο κόμβων και συνδέσεων, αλλά στην ουσία συμπυκνώνει τον λόγο, τον ομιλητή και τις πολιτικές θητείες, της εικόνας 3.3.6.



Εικόνα 3.5.2.2 – Semantic μοντέλο για το debates dataset

3.5.2.1 RDF διασύνδεση με Wikidata

Στον τομέα της αναπαράστασης γνώσης και της διασύνδεσης δεδομένων, η σύνδεση του σχήματος RDF με άλλες οντολογίες είναι ζωτικής σημασίας. Μία τέτοια εναλλακτική οντολογία 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. Κάθε άρθρο αποτελείται από δεδομένα και ζεύγη κλειδιών-τιμών και συνδέσμους προς άλλα άρθρα, σχηματίζοντας έτσι ένα σύνολο σημασιολογικά δομημένων γραφημάτων.

Η διασύνδεση των δεδομένων που έχουν καταγραφεί για τα πρακτικά του ελληνικού κοινοβουλίου με τα αντίστοιχα του wikidata, γίνεται με την χρήση ενός web crawler μέσω ενός νέου αρχείου python. Αυτός ο web crawler έχει σχεδιαστεί να περιηγείται στο διαδίκτυο και να αναζητά στον ιστότοπο της Google το όνομα του ομιλητή, με την προσθήκη του όρου "wikidata". Το Google δίνει την δυνατότητα για εύκολη αναζήτηση μέσω της σύνταξης

$$https://www.google.com/search?q="query\_here"$$

όπου αλλάζοντας το *query_here* με το *"name_speaker wikidata"* αναζητούμε για όλους τους ομιλητές που έχουν προκύψει κατά την διάρκεια μετατροπής των xml αρχείων. Έπειτα αξιολογείται 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 αποθηκεύονται για μελλοντική χρήση.

Η διαδικασία που ακολουθείται για την αποθήκευση και την εύκολη και γρήγορη αναζήτηση του σε μετέπειτα στάδιο, βασίζεται στην κατασκευή ενός νέου JSON αρχείου, στο όποιο περιέχονται ζεύγη ονομάτων και διευθύνσεων URL. Στην περίπτωση που το όνομα που αναζητήσαμε, δεν επιστρέφει αποτέλεσμα στο wikidata, το όνομα δεν τοποθετείται στον JSON αρχείο, με αποτέλεσμα μετά να μην δημιουργηθεί τριπλέτα για το ομιλητή και το wikidata_link.

Διασταυρώνοντας τα ονόματα των ομιλητών με το παραπάνω αρχείο JSON, εξασφαλίσαμε ότι είχαμε πρόσβαση στους πιο ενημερωμένους και ακριβείς συνδέσμους Wikidata για τις οντότητες στο σύνολο δεδομένων μας. Αυτή η σύνδεση μας επέτρεψε να εμπλουτίσουμε τα δεδομένα RDF μας με πρόσθετες πληροφορίες από το Wikidata, διευκολύνοντας την πιο ολοκληρωμένη και πλούσια σε περιεχόμενο αναπαράσταση γνώσης.

Τέλος και προκειμένου να επιτευχθεί όσο τον δυνατόν υψηλότερη διασύνδεση με την wikidata επιλέχθηκε να επαναληφθεί η παραπάνω διαδικασία τόσο και για τα κόμματα όσο και για τους υπουργικά αξιώματα. Έτσι δημιουργήθηκαν τριπλέτες rdf για τα κόμματα συνδέοντας

κάθε κόμμα με το αντίστοιχο url προς την βάση wikidata. Όσον αφορά τις υπουργικές θέσεις, επιλέχθηκε να συνδεθεί η κάθε θέση με το αντίστοιχο υπουργείο γενικότερα, και όχι με την αντίστοιχη συγκεκριμένη θέση.

3.5.2.2 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 Υποθέσεων, μια συλλογή αξιόπιστων νομικών και κοινοβουλευτικών πληροφοριών, είναι μια τέτοια ανεκτίμητη πηγή (https://gslegal.gov.gr/?page_id=776&sort=time). 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, οι τριπλέτες RDF μας, θα ενισχυθούν με στοιχεία και λεπτομερείς πληροφορίες σχετικά με τα κυβερνητικά πλαίσια, τις νομοθετικές διαδικασίες και τους βουλευτές. Με άλλα λόγια, έχουν εκμεταλλευτεί 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.

Επίσης χρήσιμες πληροφορίες εξάγουμε και από τον επίσημο ιστότοπο του Ελληνικού Κοινοβουλίου (https://www.hellenicparliament.gr/Vouleftes/Diatelesantes-Vouleftes-Apo-Ti-Metapolitefsi-Os-Simera/), ενισχύοντας την αναπαράσταση της γνώσης μας περιλαμβάνει με περεταίρω πληροφορίες για τους βουλευτές. Πιο συγκεκριμένα, έχουν αντληθεί 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.

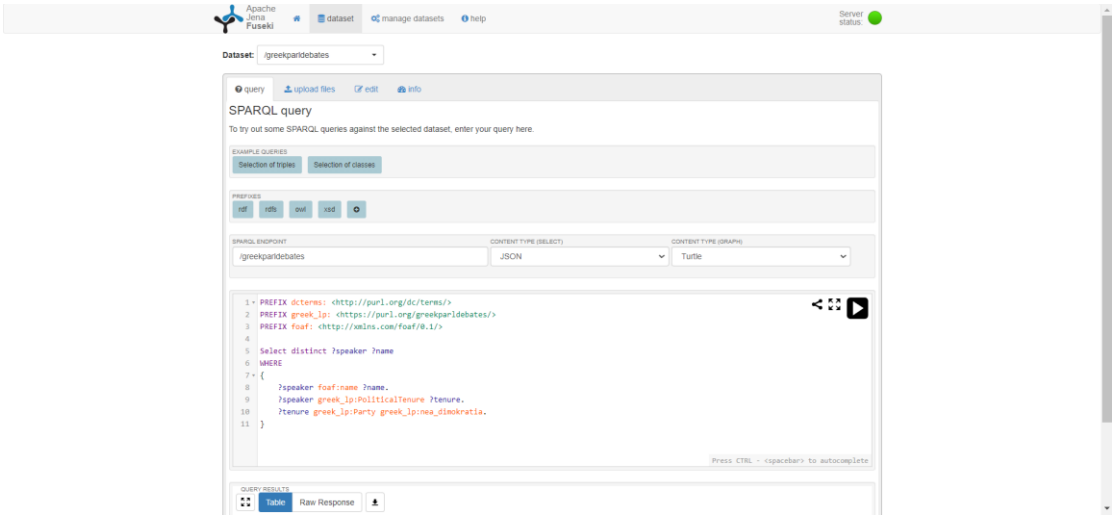
Η ενσωμάτωση αυτών των δημόσιων δεδομένων στην αναπαράσταση της γνώσης μας με τη χρήση RDF μας εντείνουν το κύρος και την αξιοπιστία. Προσφέρει πληρέστερη κατανόηση των νομοθετικών και διοικητικών διαδικασιών, τοποθετώντας τις κοινοβουλευτικές διαδικασίες στην προοπτική της ευρύτερης κυβερνητικής δομής. Επιπλέον, καθώς οι πληροφορίες από επίσημες πηγές φέρουν υψηλό επίπεδο αυθεντικότητας και αξιοπιστίας, η ενσωμάτωση αυτή ενθαρρύνει τη διαφάνεια και τη λογοδοσία.

3.5.3 Επεξεργασία RDF – SPARQL

Σε επόμενο βήμα βρίσκεται η αξιοποίηση των αρχείων RDF που παράγονται από τη διαδικασία μετατροπής XML σε RDF. Αυτά τα αρχεία RDF είναι εμπλουτισμένα με δεδομένα που αντιπροσωπεύουν κοινοβουλευτικές συνεδριάσεις, ομιλίες, ομιλητές, πολιτικές θητείες και κυβερνητικούς ρόλους και επομένως είναι έτοιμα να χρησιμοποιηθούν για περαιτέρω ανάλυση και μελέτη.

Χρησιμοποιούμε το Apache Jena Fuseki, ένα ισχυρό σύστημα βάσεων δεδομένων RDF, για να επιτρέψουμε την απλή πρόσβαση και την υποβολή ερωτημάτων στα δεδομένα RDF. Δημιουργώντας την βάση δεδομένων, αναρτούμε όλα τα αρχεία που έχουν προκύψει και είναι απαραίτητα για την εργασία. Αφού ολοκληρωθεί η διαδικασία αυτή, μπορούμε να διατυπώσουμε ακριβή ερωτήματα και να λαμβάνουμε συγκεκριμένα δεδομένα από τα τελικά συνδεδεμένα δεδομένα, χρησιμοποιώντας την SPARQL, τη γλώσσα ερωτημάτων για RDF.

Την χρονική περίοδο που γράφεται και εκπονείται η εργασία αυτή έχει δημιουργηθεί ένας server, όπου μέσω της ιστοσελίδας http://diplomatikes.dslab.ece.ntua.gr:3030/manage.html, μπορεί ο καθένας να έχει πρόσβαση στην βάση μας και να υποβάλει ερωτήματα SPARQL, επάνω στις ήδη υπάρχουσες τριπλέτες. Το περιβάλλον χρήσης φαίνεται παρακάτω όπου στο πρώτο σχήμα φαίνεται η θέση που θέτουμε ερώτημα (Εικόνα 3.5.3.1), και στο δεύτερο σχήμα είναι φανερό το αποτέλεσμα του ερωτήματος σε μορφή πίνακα (Εικόνα 3.5.3.2). Αναλυτική παρουσίαση παραδείγματος χρήσης του λογισμικού θα ακολουθήσει σε επόμενο κεφάλαιο.



Εικόνα 3.5.3.1 – Περιβάλλον Apache Jena Fuseki.

Εικόνα 3.5.3.2 – Δείγμα απάντησης στο Apache Jena Fuseki σε μορφή πίνακα.

3.6 Μετατροπή Αρχείων Akoma Ntoso σε ParlaMint Tei

Στην τελευταία αυτή παράγραφο, θα παρουσιαστεί ο τρόπος ακολουθήθηκε για την τελική δημιουργία των κοινοβουλευτικών αρχείων σε μορφή TEI ParlaMint.

Για τον σκοπό αυτό, χρησιμοποιήθηκαν ως αρχικά δεδομένα τα αρχεία xml, μορφής Akoma Ntoso, τα οποία αποτελούν ακριβές και δομημένη αναπαράσταση των πρακτικών της Βουλής που βρίσκονται στην βάση δεδομένων μας.

Η διαδικασία μετατροπής είναι ιδιαίτερα απλή και περιλαμβάνει την αντιστοίχιση των στοιχείων και των χαρακτηριστικών από το Akoma Ntoso στα αντίστοιχα στοιχεία του TEI ParlaMint, διασφαλίζοντας τη διατήρηση της ακεραιότητας και του νοήματος των δεδομένων.

Ως άξονας σε αυτή τη διαδικασία μετατροπής χρησιμεύει το XSLT. Στην παρούσα εφαρμογή, το αρχείο που περιλαμβάνει τους κανόνες για την αντιστοίχιση αντλήθηκε από το δημόσιο repository της δημιουργισας ομάδας Clarin, το akn2tei.xsl (https://github.com/clarin-eric/parla-clarin/blob/master/Examples/AkomaNtoso/akn2tei.xsl). Βασιζόμενοι σε αυτό το αρχείο, η γλώσσα XSLT, επιτρέπει την ακριβή αντιστοίχιση και μετατροπή των στοιχείων. Έτσι, τα περίπλοκα μοτίβα στο Akoma Ntoso XML εντοπίζονται και μεταφράζονται στη δομημένη μορφή του TEI ParlaMint.

Ενδεικτικά, ένα χαρακτηριστικό παράδειγμα της σύνταξης ενός αρχείου xml, που βασίζεται στην μορφή του TEI ParlMint βρίσκεται στην εικόνα 3.6.1. Όπως είναι φανερό το κείμενο αυτό είναι ίδιο με το απόσπασμα των προηγούμενων παραγράφων, και συγκεκριμένα της εικόνας 3.3.6.

```
<u xml:id="debate_2018-06-08_speech_12" who="anna-misel_asimakopoulou">
   <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 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.</seg>
   <seg>Ευχαριστώ.</seg>
</u>
<u xml:id="debate_2018-06-08_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_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 3.6.1 – Απόσπασμα από κύριο μέρος ενός αρχείου xml-tei ParlaMint.

Όσον αφορά την τελευταία εικόνα (3.6.1), είναι χαρακτηριστική η ιδιόμορφη σύνταξη του xml tei ParlaMint αποσπάσματος. Κάθε λόγος περικλύεται 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> και <seg>. Στο περιεχόμενο της ετικέτας <note> βρίσκεται αυτόυσιο το όνομα του ομιλητή, και μέσα στην ετικέτα <seg> περικλύονται τα τμήματα του λόγου.

3.7 Σχολιασμός περί Σφαλμάτων

Γενικά, η ομαλή μετατροπή εγγράφων από μια μορφή σε μια άλλη είναι απαραίτητη στον σύγχρονο κόσμο του διαδικτύου. Ωστόσο, υπάρχουν αρκετές δυσκολίες, που εμποδίζουν

μια αυτοματοποιημένη διαδικασία να εκτελεί αυτή τη μετατροπή, κάποιες από τις οποίες συναντήθηκαν και εδώ.

Στην παρούσα εργασία, όπως είδαμε και παραπάνω, σε πρώτο στάδιο, καλούμασταν να κατασκευάσουμε ένα σύστημα στο οποίο θα εισάγονται αρχεία word/pdf και θα τα μετατρέπονται σε αρχεία xml, και συγκεκριμένα με μορφή Akoma Ntoso. Η πολυπλοκότητα ενός τέτοιου συστήματος σε συνδυασμό με τα ιδιόμορφα εργαλεία που χρησιμοποιούμε για την δημιουργία του, δεν διευκολύνει την δημιουργία ενός αψεγάδιαστου και αλάνθαστου συστήματος.

Στη συγκεκριμένη εργασία, καταφέραμε ένα ποσοστό επιτυχίας μετατροπής αρχείων που αγγίζει το *96%, καθώς μετατράπηκαν τα 5471 από τα 5688 αρχικά αρχεία*. 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 και αυστηρού μοντέλου το οποίο αναγνωρίζει και αντιστοιχίζει τις πληροφορίες των αρχικών κειμένων που εισάγονται. Σε αυτά τα λίγα ελλαττωματικά αρχεία, παρατηρούμε πως ο κειμενογράφος έχει ξεφύγει από τους βασικούς συντακτικούς κανόνες του μοντέλου μας, 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 τη διαδικασία μετατροπής. Η παρερμηνεία μπορεί να προκύψει από ανορθόγραφες λέξεις, ακατάλληλη στίξη ή ακανόνιστη μορφοποίηση.

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

Τέλος, αξίζει να σχολιάσουμε πως η μετατροπή των αρχείων από Akoma Ntoso αρχεία σε μορφή ParlaMint tei πραγματοποιήθηκε χωρίς κανένα εμπόδιο. Πιο συγκεκριμένα, η τυποποιημένη δομή των αρχείων XML/Akoma Ntoso 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. 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, όπως παραπάνω.

4

Μελέτη Περιπτώσεων και Στατιστικά Δεδομένα

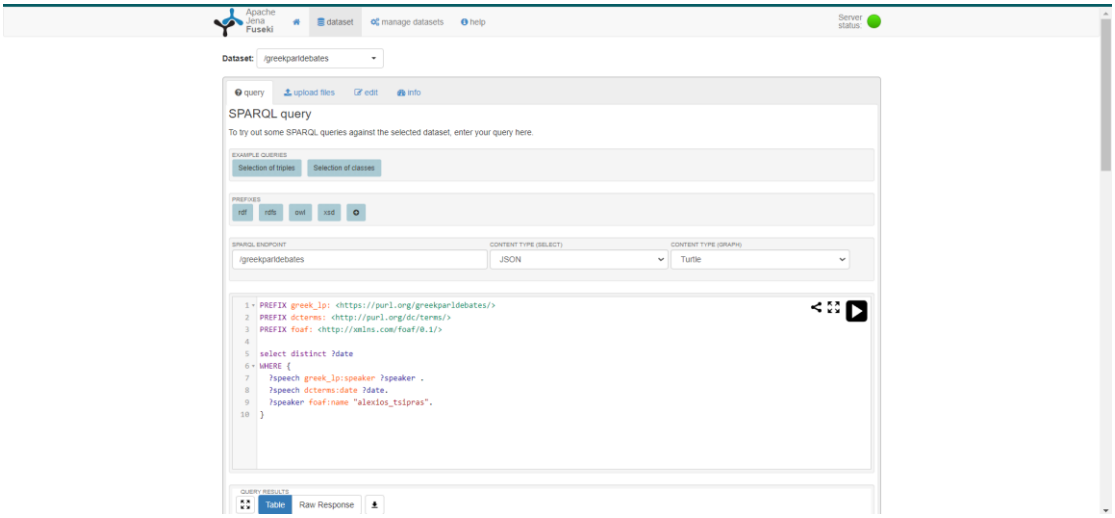
Μετά από μια εκτενή ανάλυση των θεωρητικών και πρακτικών πτυχών της επεξεργασίας των κοινοβουλευτικών διαδικασιών, σειρά έχει μια ειδική ενότητα για την παρουσίαση χρήσιμων και ενδιαφέρων περιπτώσεων τόσο πάνω στα αρχεία xml, όσο και πάνω στις πληροφορίες που προσφέρουν οι τριπλέτες RDF. Η σημασία αυτής της ενότητας δίνεται προκειμένου να δώσει στον αναγνώστη μια πιο απλοϊκή και κατανοητή ανάλυση των τελικών αποτελεσμάτων του συστήματος.

Τέλος, θα παρουσιαστούν κάποια χρήσιμα και ενδιαφέροντα στατιστικά που προέκυψαν.

4.1 Περίπτωση μελέτης RDF

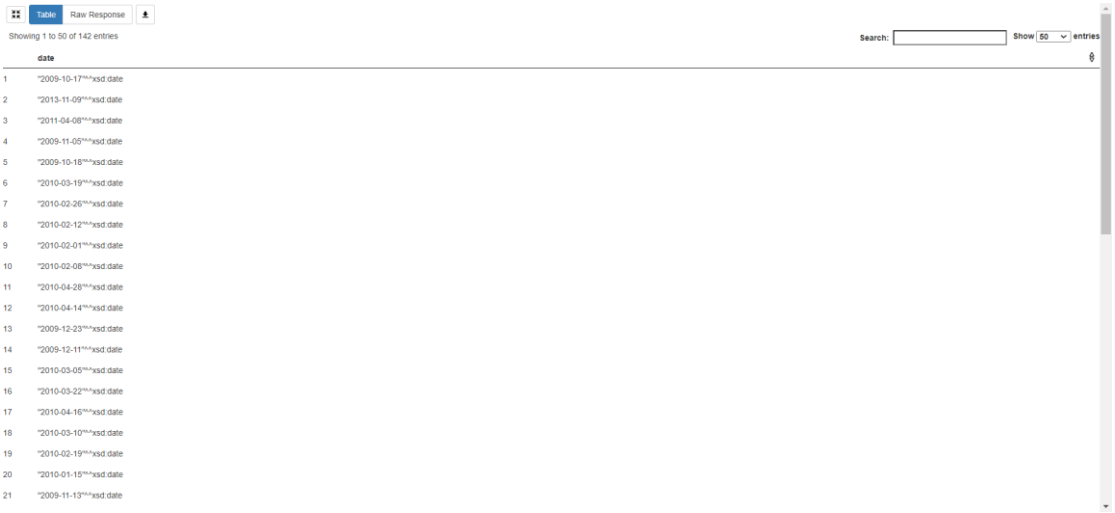
Έχοντας καταλάβει τις βασικές έννοιες και τους κύριους πυλώνες λειτουργίας του συστήματος μας, είμαστε σε θέση να παρουσιάσουμε κάποια ενδεικτικά παράδειγμα στο τελικό ‘προϊόν’ της εργασίας. Με τον όρο ‘προϊόν’ χαρακτηρίζουμε το σύνολο των τριπλέτων RDF που έχουν προκύψει με την πάροδο εκτέλεσης όλου του συστήματος και έχουν αναρτηθεί στο λογισμικό Apache Fuseki.

Έστω, για αρχή, ότι επιθυμούμε να εντοπίσουμε πόσες φορές μίλησε ένας πολιτικός μέσα στο ελληνικό κοινοβούλιο, αναμένοντας και τις αντίστοιχες ημερομηνίες. Κάνοντας την σύμβαση ότι το χρονικό πεδίο ενδιαφέροντος είναι όλο αυτό που μελετάμε (αρχές δεκαετίας 1990 – 2022) και επιλέγοντας ενδεικτικά τον βουλευτή “Αλέξη Τσίπρα” γράφουμε κατάλληλα το ερώτημα που θέλουμε στην γλώσσα ερωτημάτων SPARQL. Όπως φαίνεται και παρακάτω στο δομημένο και ολοκληρωμένο ερώτημα, έχει επιλεχθεί η αναζήτηση του ομιλητή (speaker) με όνομα “alexios_tsipras”.



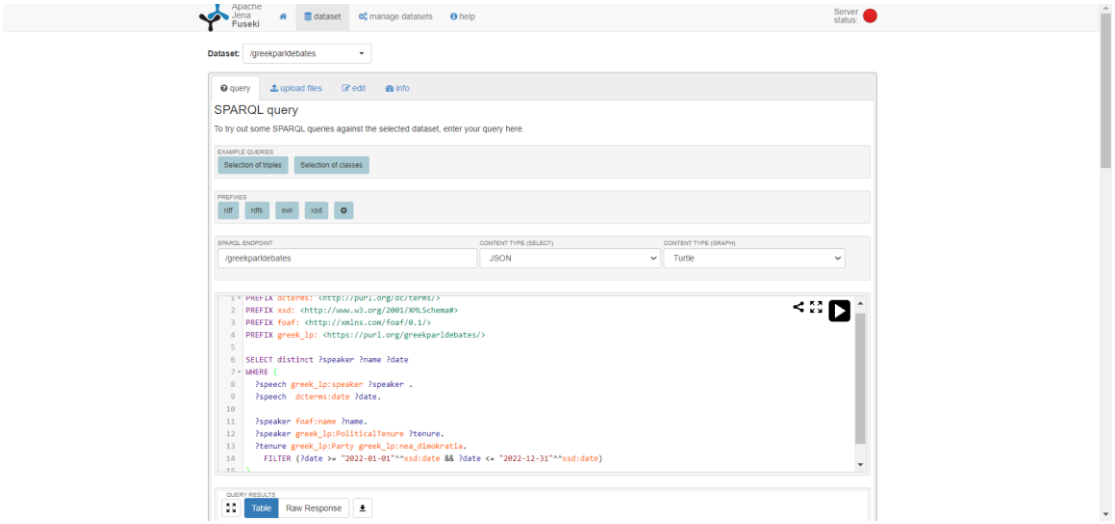
Εικόνα 4.1.1 – Παράδειγμα χρήσης – ερώτημα 1 με «Αλέξιος Τσίπρα»

Έπειτα και αφού τρέξει το ερώτημα εμφανίζεται στο κάτω μέρος της σελίδας στο πλαίσιο που τιτλοφορείται “QUERY RESULTS”, το αποτέλεσμα που προκύπτει από το ερώτημα. Στο συγκεκριμένο ερώτημα, με αντικείμενο μελέτης τον Αλέξη Τσίπρα, το αποτέλεσμα είναι ένας πίνακας με όλες τις ημερομηνίες που έχει μιλήσει. Να σημειωθεί το γεγονός πως εφόσον έχουμε χρησιμοποιήσει τον όρο “distinct” στο ερώτημα, αναμένουμε το αποτέλεσμα να μην περιέχει διπλότυπες ημερομηνίες.



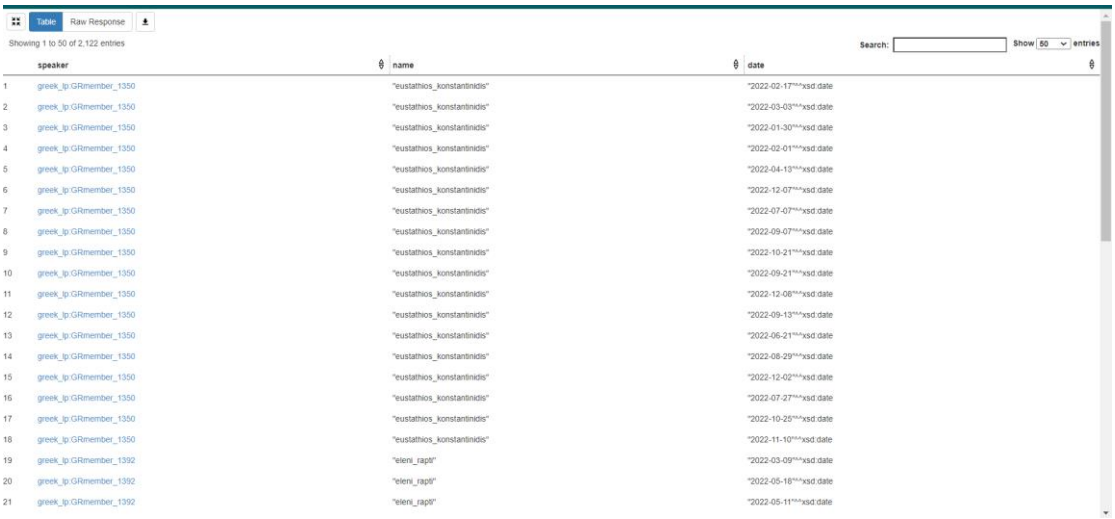
Εικόνα 4.1.2 – Παράδειγμα χρήσης – αποτέλεσμα στο ερώτημα 1 με «Αλέξιος Τσίπρα»

Σε ένα δεύτερο παράδειγμα, θα μπορούσε να τεθεί 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 αποτυπώνεται το ερώτημα όπου χαρακτηριστικά φαίνεται ο ενδιαφερόμενος όρος “nea_dimokratia”, αλλά και το χρονικό διάστημα.



Εικόνα 4.1.3 – Παράδειγμα χρήσης – ερώτημα 2 με βουλευτές Νέας Δημοκρατίας

Το αποτέλεσμα του ερωτήματος φαίνεται στην εικόνα 4.2.4, όπου φαίνονται 2.122 μοναδικές εγγραφές, όπου φαίνεται το μοναδικό αναγνωριστικό κάθε ομιλητή (στήλη *speaker*), το όνομα του όπως είναι αποθηκευμένο στην βάση δεδομένων Apache Fuseki (στήλη *name*) και τέλος η κάθε ημερομηνία (στήλη *date*).



Εικόνα 4.1.4 – Παράδειγμα χρήσης – αποτέλεσμα στο ερώτημα 2 με βουλευτές της Νέας Δημοκρατίας

4.2 Περίπτωση μελέτης XML – LDA Topic Modeling

Η ανάλυση των νομοθετικών συζητήσεων αποτελεί ακρογωνιαίο λίθο για την κατανόηση του πολιτικού περιβάλλοντος και των πολιτικών τάσεων στον τομέα της

διακυβέρνησης με βάση τα δεδομένα. Στο πλαίσιο αυτό, ο τρόπος με τον οποίο ερμηνεύονται και κατανοούνται τα δεδομένα κειμένου έχει μετασχηματιστεί πλήρως με την ενσωμάτωση νέων μεθοδολογιών, όπως η Latent Dirichlet Allocation (LDA).

Στο πλαίσιο της διαδικασίας με τίτλο θεματική μοντελοποίηση (topic modelling), συγκεντρώνονται αρχικά μεγάλες ποσότητες δεδομένων απλού κειμένου από επίσημες συζητήσεις. Στην συγκεκριμένη εργασία, επιλέχτηκε να αντληθεί το αποκλειστικά το κείμενο το διαλόγων από τα αρχεία xml – Akoma Ntoso, μιας και από αυτά τα δομημένα έγγραφα η διαδικασία εξαγωγής του κειμένου είναι πια απλή και πιο ασφαλής.

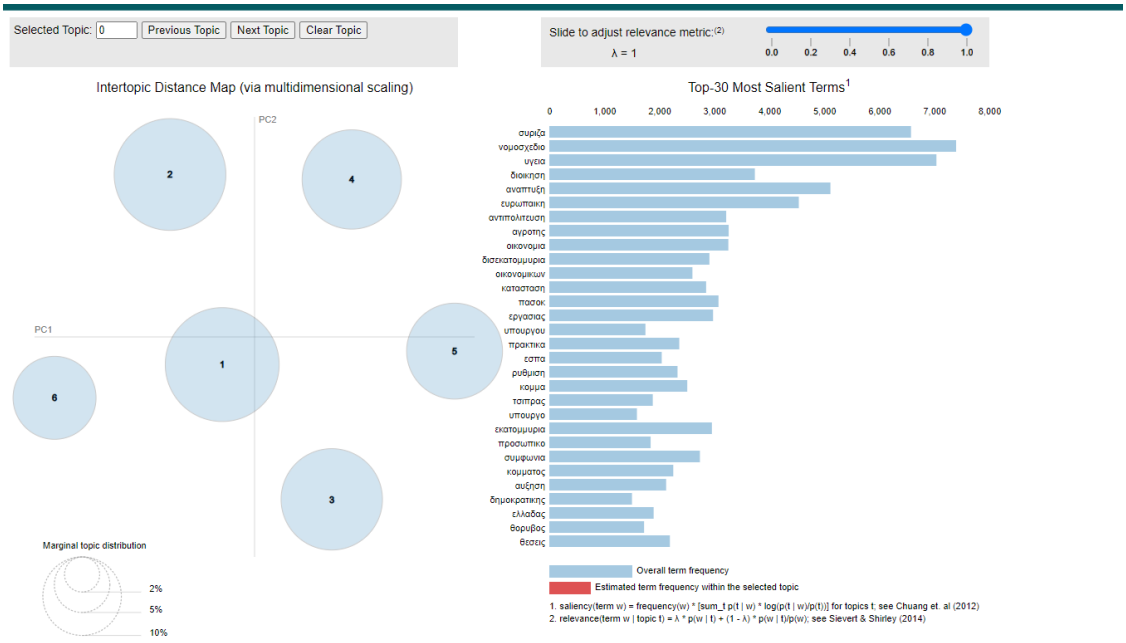
Συνήθως, το αδόμητο κειμενικό υλικό είναι περίπλοκο και ιδιαίτερα εκτενές. Σε αυτή την κατάσταση, ένα ισχυρό μοντέλο – το Latent Dirichlet allocation (LDA) – αναδεικνύεται σε βασικό εργαλείο για την επεξεργασία και οργάνωση του κειμένου. Πιο συγκεκριμένα, το παραγωγικό στατιστικό μοντέλο της LDA, έχει την δυνατότητα να βρίσκει τις λανθάνουσες θεματικές δομές σε μεγάλα σύνολα δεδομένων κειμένου. Ο στόχος της χρήσης της LDA για την ανάλυση των κυβερνητικών συζητήσεων είναι να συμπυκνώσει συνομιλίες ετών σε διαχειρίσιμα θέματα, ρίχνοντας φως στα επαναλαμβανόμενα θέματα και ζητήματα κάθε έτους.

Η μοντελοποίηση θεμάτων, μια μέθοδος που χρησιμοποιείται από την LDA για να ταξινομήσει το κείμενο και να βρει λέξεις και φράσεις που συνυπάρχουν τακτικά, είναι το κύριο σημείο αυτού του υπό-συστήματος. Οι κοινοί όροι κατηγοριοποιούνται σε θέματα, καθένα από τα οποία αντιπροσωπεύει ένα διαφορετικό θέμα 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.

Παρόλο που η LDA παρέχει βαθιές γνώσεις, εξακολουθούν να υπάρχουν ζητήματα όπως ο θόρυβος των δεδομένων κειμένου και η ερμηνευσιμότητα των θεμάτων. Προκειμένου να ξεπεραστεί αυτή η 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.

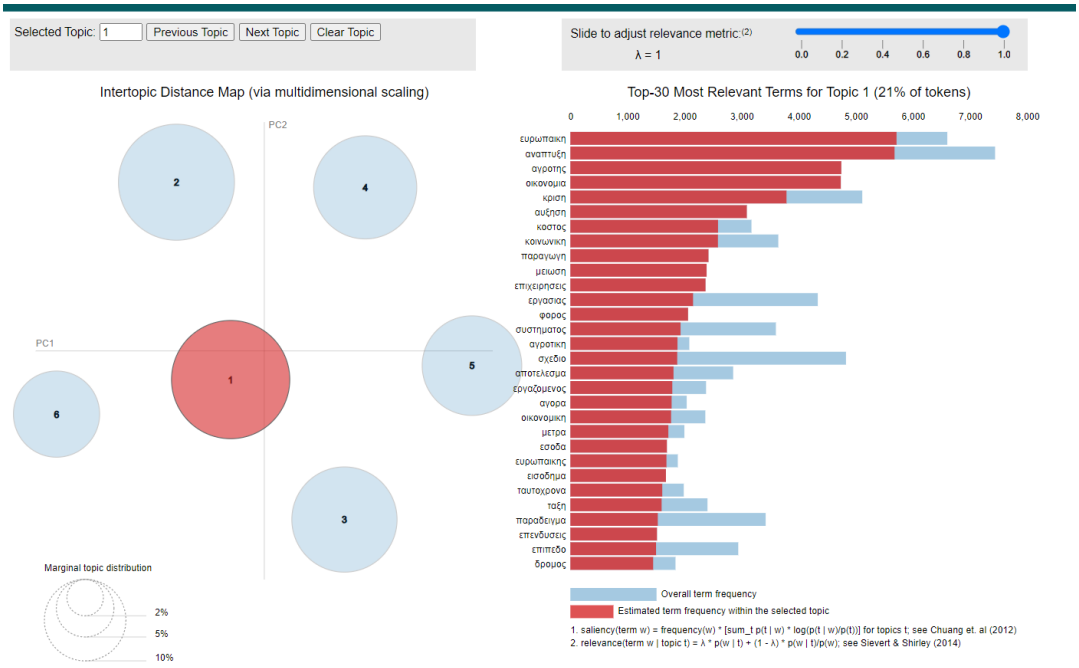
Το αποτέλεσμα του μοντέλο LDA αξιοποιείται καταλλήλως από την βιβλιοθήκη της Python, “pyldavis”, η οποία τα οπτικοποιεί και τα αποτυπώνει όπως φαίνεται στην εικόνα 4.2.1. Στο πρακτικό επίπεδο, στο δεξί μέρος φαίνεται μία λίστα με τους τριάντα (30) πιο «σημαντικούς» όρους, με την αντίστοιχη συχνότητα εμφάνισης. Στο αριστερά, φαίνεται ένα διάγραμμα με τα θέματα που έχουν προκύψει και την μεταξύ τους συσχέτιση ανάλογα με την απόσταση τους. Τέλος, με την επιλογή ενός εκ των θεμάτων, προσαρμόζεται καταλλήλως η λίστα και η συχνότητα των λέξεων.

Ενδεικτικά να αναφέρουμε πως 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 4.2.1 – Κύρια οθόνη Topic Modeling αποτελέσματος με LDA για έτος 2016

Εστιάζοντας παραπάνω για το έτος 2016, θα προσπαθήσουμε να ‘μαντέψουμε’ μερικά από τα θέματα που συζητήθηκαν στο ελληνικό κοινοβούλιο. Τυχαία επιλέγουμε το θέμα με νούμερο 1 (topic 1) και παρατηρούμε πως στην κορυφή βρίσκονται όροι που αφορούν την οικονομία, την ευρωπαϊκή, μια κρίση, αλλά και παράγοντες όπως η εργασία και η αγροτική ζωή (εικόνα 4.2.2). Συνδέοντας όλα αυτά μαζί 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 όποια μπορεί εύκολα να επιβεβαιωθεί ή να διαψευστεί μέσω μιας απλής αναζήτησης και έρευνας στο διαδίκτυο.



Εικόνα 4.2.2 – Οθόνη Topic Modeling αποτελέσματος με LDA για έτος 2016 με επιλεγμένο το θέμα 1

Αναλύοντας και επεξεργάζοντας περαιτέρω τα αποτελέσματα που προέκυψαν από το μοντέλου Latent Dirichlet allocation (LDA) συγκεντρώσαμε σε έναν ενιαίο πίνακα τους δέκα πιο σημαντικούς όρους των συζητήσεων συνολικά για κάθε έτος. Αυτοί οι όροι προσδιορίστηκαν με βάση τη σημασία και τη συνάφεια τους στο σύνολο δεδομένων, αναζητώντας κάθε φορά την συχνότητα εμφάνισης 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 εξάγονται, κάποια αναμενόμενα και κάποια μη, συμπεράσματα. Αρχικά, 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 μαστίζουν την χώρα από το 2008 και έπειτα.

Πίνακας 4.2.1 – Πρώτοι δέκα σημαντικοί όροι για κάθε έτος (από μοντέλο LDA)

Top 10 Most Salient Terms Overall (per year)

Year:1989	**Year:1999**	**Year:2008**	**Year:2017**
πασοκ: 2806	νομοσχεδιο: 5282	πασοκ: 11661	νομοσχεδιο: 5728
νομοσχεδιο: 1073	αναπτυξη: 4413	νομοσχεδιο: 9415	υγεια: 5147
συμβουλιο: 855	υγεια: 3897	αναπτυξη: 7828	συριζα: 4125
δικη: 809	ευρωπαικη: 3699	ευρωπαικη: 5575	αναπτυξη: 3935
δικαιωμα: 794	συμβαση: 3399	υγεια: 5383	εργασιας: 2942
κοινοτητα: 728	παιδεια: 3298	κοινωνια: 4836	πασοκ: 2523
συμβουλιου: 713	διοικηση: 2875	αντιπολιτευση: 4460	αντιπολιτευση: 2488
διαταξη: 694	ρυθμιση: 2873	κριση: 4379	ευρωπαικη: 2440
επιτροπης: 690	πασοκ: 2857	σχεδιο: 4317	οικονομια: 2172
πλευρα: 671	δραχμας: 2748	εκπαιδευση: 4146	ρυθμιση: 2107
Year:1990	**Year:2000**	**Year:2009**	**Year:2018**
πασοκ: 7641	νομοσχεδιο: 4082	πασοκ: 8367	συριζα: 13111
νομοσχεδιο: 5541	αναπτυξη: 2941	νομοσχεδιο: 4925	νδ: 5409
αυξηση: 3615	πασοκ: 2541	αναπτυξη: 4889	νομοσχεδιο: 4535
κοινοτητα: 3602	διοικηση: 2385	κριση: 4889	αναπτυξη: 3690
αθηνα: 3288	ευρωπαικη: 2153	υγεια: 3561	υγεια: 2828
οικονομια: 3255	στηριξη: 1814	οικονομια: 3166	κεντρωων: 2445
δικαιωμα: 3206	εργασιας: 1797	ευρωπαικη: 3089	εργασιας: 2407
διαταξη: 2840	οικονομια: 1793	κατασταση: 2792	συμφωνια: 2382
οικονομικων: 2840	εκπαιδευση: 1785	εργασιας: 2620	σχεδιο: 2313
αναπτυξη: 2812	ρυθμιση: 1753	σχεδιο: 2615	ευρωπαικη: 2308
Year:1991	**Year:2001**	**Year:2010**	**Year:2019**
νομοσχεδιο: 6247	νομοσχεδιο: 6970	νομοσχεδιο: 10264	συριζα: 20572
πασοκ: 5008	υγεια: 5021	πασοκ: 9184	νομοσχεδιο: 5735
κοινοτητα: 4410	αναπτυξη: 4437	αναπτυξη: 6681	αναπτυξη: 4520
αθηνα: 4176	ευρωπαικη: 3965	υγεια: 5513	αθηνων: 4469
διαταξη: 3488	πασοκ: 3496	κριση: 5161	υγεια: 4358
δικη: 3233	διαταξη: 3149	οικονομια: 4182	αλλαγης: 3608
διοικηση: 3034	διοικηση: 2907	ευρωπαικη: 4047	νδ: 3557
δικαιωμα: 3020	ελεγχος: 2669	κοινωνια: 3883	συμφωνια: 3537
βουλευτων: 2833	γεωργιας: 2596	σχεδιο: 3849	κκε: 3457
παιδεια: 2820	ρυθμιση: 2541	κατασταση: 3822	κωνσταντινος: 3371
Year:1992	**Year:2002**	**Year:2011**	**Year:2020**
νομοσχεδιο: 7216	νομοσχεδιο: 3848	νομοσχεδιο: 8331	συριζα: 25738
κοινοτητα: 6299	αναπτυξη: 2707	πασοκ: 8141	υγεια: 10390
υγεια: 5775	ευρωπαικη: 2641	αναπτυξη: 6913	αλλαγης: 9414

πασοκ: 4733	διοικηση: 2421	υγεια: 6310	νομοσχεδιο: 9395
διαταξη: 3646	πασοκ: 2290	κριση: 5178	αθηνων: 8489
γεωργιας: 3496	υγεια: 1970	ευρωπαικη: 4501	κινημα: 8174
αναπτυξη: 3345	ελεγχος: 1854	κοινωνια: 4197	λυση: 8122
δικαιωμα: 3316	διαταξη: 1831	οικονομια: 4101	κκε: 6459
αυξηση: 3301	δικη: 1724	κατασταση: 3960	αναπτυξη: 5596
διοικηση: 3211	εργασιας: 1716	σχεδιο: 3439	κριση: 5023
Year:1993	**Year:2003**	**Year:2012**	**Year:2021**
νομοσχεδιο: 5428	νομοσχεδιο: 6007	αναπτυξη: 4997	συριζα: 26295
πασοκ: 4611	υγεια: 4812	νομοσχεδιο: 4679	νομοσχεδιο: 10080
κοινοτητα: 3893	αναπτυξη: 4128	υγεια: 4262	υγεια: 9285
διαταξη: 3649	ευρωπαικη: 3572	κριση: 3376	αλλαγης: 8982
δικαιωμα: 2798	δημοσιας: 3358	πασοκ: 3210	κινημα: 7746
δικη: 2454	δημος: 3357	ευρωπαικη: 2910	λυση: 7722
υγεια: 2329	διοικηση: 2877	οικονομικων: 2844	αναπτυξη: 6670
συμβουλιο: 2314	εργασιας: 2833	εργασιας: 2624	κκε: 6034
εθνικης: 2283	ελεγχος: 2477	κατασταση: 2512	εργασιας: 5950
εθνικη: 2234	πασοκ: 2467	κοινωνια: 2457	αθηνων: 5296
Year:1994	**Year:2004**	**Year:2013**	**Year:2022**
νομοσχεδιο: 4896	αναπτυξη: 5524	νομοσχεδιο: 7828	συριζα: 18441
κοινοτητα: 4104	νομοσχεδιο: 5186	υγεια: 5840	υγεια: 8606
πασοκ: 3582	πασοκ: 4787	αναπτυξη: 5752	νομοσχεδιο: 7915
υγεια: 2618	υγεια: 3437	συριζα: 4552	αλλαγης: 7858
διαταξη: 2481	διοικηση: 3102	εργασιας: 4422	λυση: 6849
αναπτυξη: 2308	ευρωπαικη: 2815	οικονομικων: 4060	αθηνων: 6702
ρυθμιση: 2222	αντιπολιτευση: 2397	κριση: 3849	κινημα: 6459
δικη: 2125	ρυθμιση: 2357	σχεδιο: 3814	ενεργεια: 6115
εθνικη: 2046	οικονομια: 2146	ευρωπαικη: 3584	κκε: 5194
μερος: 2033	απασχοληση: 2129	κοινωνια: 3159	αναπτυξη: 5129
Year:1996	**Year:2005**	**Year:2014**	
αναπτυξη: 1551	πασοκ: 9822	νομοσχεδιο: 8100	
ευρωπαικη: 1173	νομοσχεδιο: 9550	υγεια: 6462	
πασοκ: 1146	αναπτυξη: 8376	αναπτυξη: 5841	
νομοσχεδιο: 1037	ευρωπαικη: 6355	συριζα: 4288	
αυξηση: 940	υγεια: 5296	ευρωπαικη: 3568	
εθνικη: 929	αντιπολιτευση: 3858	εργασιας: 3364	
δραχμας: 880	οικονομια: 3545	κριση: 3225	
οικονομικων: 860	αγορα: 3360	ρυθμιση: 2961	
εθνικης: 841	εκπαιδευση: 3352	ενεργεια: 2888	
διαταξη: 827	σχεδιο: 3321	σχεδιο: 2852	

Year:1997	**Year:2006**	**Year:2015**
νομοσχεδιο: 8581	πασοκ: 8896	νομοσχεδιο: 5220
υγεια: 5904	αναπτυξη: 6814	συριζα: 4416
αναπτυξη: 5593	νομοσχεδιο: 6667	ευρωπαικη: 2645
κοινοτητα: 5090	υγεια: 3629	υγεια: 2419
δημος: 4588	ευρωπαικη: 3573	κριση: 2375
διοικηση: 4442	εκπαιδευση: 3351	αναπτυξη: 2257
διαταξη: 4094	αυξηση: 3081	συμφωνια: 2218
οικονομικων: 4015	αντιπολιτευση: 3020	οικονομια: 2210
γεωργιας: 4006	δημος: 2904	κοινωνια: 2041
μερος: 3975	σχεδιο: 2728	παιδεια: 2027
Year:1998	**Year:2007**	**Year:2016**
νομοσχεδιο: 2580	πασοκ: 8668	νομοσχεδιο: 6650
αναπτυξη: 2369	αναπτυξη: 5881	συριζα: 5927
υγεια: 2264	νομοσχεδιο: 5648	αναπτυξη: 5261
πασοκ: 2080	υγεια: 5432	υγεια: 4621
παιδεια: 2041	ευρωπαικη: 3142	οικονομια: 3414
δημοσιας: 2030	αντιπολιτευση: 3123	ευρωπαικη: 3375
δραχμας: 2018	παιδεια: 2826	κριση: 2963
ευρωπαικη: 1986	αυξηση: 2654	πασοκ: 2915
εργασιας: 1964	σχεδιο: 2562	σχεδιο: 2894
διοικηση: 1909	διοικηση: 2512	αντιπολιτευση: 2894

4.3 Στατιστικά Δεδομένα

Σε αυτή την παράγραφο της διπλωματικής εργασίας θα παρατεθούν κάποιοι πίνακες στους οποίους αποτυπώνονται στατιστικά στοιχεία.

Στον πρώτο πίνακα, ΠΙΝΑΚΑΣ 4.3.1, καταγράφονται στατιστικά για το πλήθος των αρχικών αρχείων της βάσης δεδομένων κατάφεραν να μετατραπούν σε αρχεία xml, μορφής Akoma Ntoso, χωρισμένα ανά έτος. Ιδιαίτερης σημασίας είναι το γεγονός ότι σχεδόν σε όλες τις περιπτώσεις το ποσοστό επιτυχίας ξεπερνάει το φράγμα του 90%, εκτός από το έτος 1995 που δεν υπάρχει κανένα αρχείο. Φυσικά στο τέλος φαίνεται και το συνολικό ποσοστό για όλα τα έτη αθροιστικά.

ΠΙΝΑΚΑΣ 4.3.1 – Στατιστικά ανά έτος που έγιναν xml

Έτος	#εγιναν xml	#αποτυχίας	Ποσοστό επιτυχίας %
1989	45	3	94
1990	162	11	94
1991	169	14	92
1992	179	26	87
1993	150	11	93
1994	131	3	98
1995	*0*	*0*	*0*
1996	46	4	92
1997	191	5	97
1998	91	3	97
1999	171	9	95
2000	148	17	90
2001	196	13	94
2002	150	17	90
2003	165	9	95
2004	141	8	95
2005	194	14	93
2006	187	1	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	207	0	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	186	2	99
ΣΥΝΟΛΙΚΑ	**5471**	**217**	**96**

Στο ίδιο μήκος κύματος με παραπάνω, στον ΠΙΝΑΚΑΣ 4.3.2, καταγράφονται παρόμοια στατιστικά με την διαφορά ότι σε αυτήν την περίπτωση έχει γίνει διαχωρισμός 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 4.3.2 – Στατιστικά ανά Κοινοβουλευτική Περίοδο που έγιναν xml

Αριθμός Περιόδου*	#εγιναν xml	#αποτυχίας	Ποσοστό επιτυχίας %
Ε' ΠΕΡΙΟΔΟΣ	33	2	94
Ζ' ΠΕΡΙΟΔΟΣ	593	56	91
Η' ΠΕΡΙΟΔΟΣ	163	9	95
Θ΄ ΠΕΡΙΟΔΟΣ	539	126	81
Ι' ΠΕΡΙΟΔΟΣ	641	87	88
ΙΑ' ΠΕΡΙΟΔΟΣ	628	26	96
ΙΒ' ΠΕΡΙΟΔΟΣ	372	14	96
ΙΓ΄ ΠΕΡΙΟΔΟΣ	552	6	99
ΙΔ΄ ΠΕΡΙΟΔΟΣ	3	0	100
ΙΕ΄ ΠΕΡΙΟΔΟΣ	479	7	99
ΙΖ΄ ΠΕΡΙΟΔΟΣ	685	9	99
ΙΗ΄ ΠΕΡΙΟΔΟΣ	644	10	98
ΙΣΤ΄ ΠΕΡΙΟΔΟΣ	92	3	97
ΣΤ' ΠΕΡΙΟΔΟΣ	47	4	92
ΣΥΝΟΛΙΚΑ	*5464*	*366*	94

* (ΠΡΟΕΔΡΕΥΟΜΕΝΗΣ ΚΟΙΝΟΒΟΥΛΕΥΤΙΚΗΣ ΔΗΜΟΚΡΑΤΙΑΣ)

5

Επίλογος

Ολοκληρώνοντας θα ήταν απαραίτητο σε αυτή την τελευταία ενότητα να συνοψίσουμε τις βασικές πτυχές της εργασίας. Καθ' όλη τη διάρκεια αυτής της έρευνας, το κύριο μέλημά μας ήταν η μετατροπή των συζητήσεων της Βουλής των Ελλήνων σε μορφές XML και RDF, υιοθετώντας τις αρχές των ανοικτών συνδεδεμένων δεδομένων. Στόχος ήταν να καταστήσουμε τον εκτεταμένο πλούτο νομοθετικών πληροφοριών προσβάσιμο και αξιοποιήσιμο σε ένα διασυνδεδεμένο ψηφιακό περιβάλλον. Αντιμετωπίζοντας σχολαστικά τις προκλήσεις που συνδέονται με αυτόν τον μετασχηματισμό, δομήσαμε με επιτυχία τις συζητήσεις, εξασφαλίζοντας τη συμβατότητά τους με τον Σημασιολογικό Ιστό.

5.1 Σύνοψη και συμπεράσματα

Υπό το πρίσμα των συζητήσεων στο Ελληνικό Κοινοβούλιο, η παρούσα διατριβή αποτελεί καίριο σημείο καμπής στους τομείς του ψηφιακού μετασχηματισμού και των ανοικτών συνδεδεμένων δεδομένων. Ο κύριος στόχος αυτής της εργασίας ήταν η μετατροπή των κοινοβουλευτικών συζητήσεων σε μορφές XML και RDF, μετατρέποντας ένα μεγάλο πλήθος κειμένων σε δεδομένα που θα μπορούσαν να προσπελαστούν και να χρησιμοποιηθούν με άλλα συστήματα. Καταφέραμε να επιτύχουμε αυτόν τον στόχο με ενδελεχή έρευνα, προσεκτική ανάλυση και χρήση τεχνολογιών αιχμής του Σημασιολογικού ιστού, γεγονός που αποτελεί ένα βήμα προς τη σωστή κατεύθυνση για τη βελτίωση της ανοικτότητας, της χρηστικότητας και της προσβασιμότητας των νομοθετικών διαδικασιών.

Η αποτελεσματική μετατροπή των συζητήσεων σε ανοικτά δια-συνδεδεμένα δεδομένα είναι ένα από τα σημαντικότερα θετικά αποτελέσματα της παρούσας εργασίας. Το επίτευγμα αυτό δημιουργεί νέες ευκαιρίες για τους μελετητές, τους φορείς λήψης αποφάσεων και το ευρύ κοινό 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 αρχεία XML και τριπλέτες RDF.

+++

Τέλος, είναι σημαντικό να αναγνωρίσουμε πως το ταξίδι της ανάλυσης, μετατροπής και επεξεργασίας των πρακτικών του ελληνικού κοινοβούλιο ήταν τόσο σύντομο και περιεκτικό, αλλά ταυτόχρονα τόσο κρίσιμο και χρήσιμο για την ψηφιακή εκδημοκράτηση του δημόσιου λόγου.

5.2 Εμπόδια – Προβλήματα

Καθ' όλη τη διάρκεια της διπλωματικής μου εργασίας, αντιμετώπισα 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 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. Όπως είδαμε και προηγουμένως, λόγω του γεγονότος ότι τα πρακτικά του Ελληνικού Κοινοβουλίου έχουν συγγραφεί από άνθρωπο, τέτοια «περίεργα» 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 ακριβής ταυτοποίησης και αναγνώρισης ονομάτων σε ένα τέτοιο μεγάλο σύνολο δεδομένων.

5.3 Μελλοντικές επεκτάσεις

Στο πλαίσιο της επέκτασης της παρούσας διπλωματικής εργασίας, οι δυνατότητα αξιοποίηση των δεδομένων εγγυείται σε πολλούς τομείς έρευνας και ανάπτυξης. Τα αυστηρά δομημένα αρχεία xml και RDF μπορούν να υπερεκτιμηθούν στην ψηφιακή εποχή, μιας και η εξαγωγή πληροφοριών είναι μια πιο εύκολη διαδικασία.

Για αρχή, μια πιθανή 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 της διπλωματικής εργασίας ώστε να συμπεριλάβει τη μετατροπή και δημοσίευση των κοινοβουλευτικών συζητήσεων σε πραγματικό ή σχεδόν πραγματικό χρόνο θα μπορούσε να αποτελέσει μια πρωτοποριακή εξέλιξη. Η εφαρμογή αυτοματοποιημένων συστημάτων που θα μπορούν να επεξεργάζονται και να μετατρέπουν τις συζητήσεις κατά τη διάρκειά τους θα παρείχε ένα ενημερωμένο και συνεχώς εξελισσόμενο σύνολο δεδομένων. Αυτή η προσέγγιση σε πραγματικό χρόνο θα μπορούσε να ανοίξει πόρτες σε καινοτόμες εφαρμογές, όπως η ανάλυση ζωντανών συζητήσεων, η παρακολούθηση συναισθημάτων και η παρακολούθηση της κοινής γνώμης, μεταμορφώνοντας τον τρόπο με τον οποίο οι πολίτες ασχολούνται με τις κοινοβουλευτικές διαδικασίες σε πραγματικό χρόνο.

Ένας άλλος συναρπαστικός δρόμος για επέκταση 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 η ελληνική νομοθεσία, θα μπορούσε να δημιουργήσει ένα ολοκληρωμένο οικοσύστημα συνδεδεμένων δεδομένων της ελληνικής διακυβέρνησης. Αυτή η διεπιστημονική προσέγγιση θα επέτρεπε στους πολίτες να διεξάγουν πιο ολιστικές αναλύσεις, αποκτώντας βαθύτερες γνώσεις για τις περίπλοκες σχέσεις μεταξύ των νομοθετικών αποφάσεων με τους διαφορετικούς τομείς.

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 συναρπαστική επέκταση. Η ανάπτυξη μοντέλων τεχνητής νοημοσύνης για την εξαγωγή πολύτιμων πληροφοριών, την πρόβλεψη νομοθετικών τάσεων ή ακόμη και την αυτοματοποίηση ορισμένων πτυχών της πολιτικής ανάλυσης θα μπορούσε να φέρει επανάσταση στον τομέα, καθιστώντας τη νομοθετική διαδικασία πιο αποτελεσματική και ενημερωμένη.

Αυτές οι πιθανές επεκτάσεις όχι μόνο βασίζονται στα θεμέλια που έθεσε η παρούσα διπλωματική εργασία, αλλά ανοίγουν επίσης 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] *Akoma ntoso | Akoma ntoso site*. (n.d.). http://www.akomantoso.org/

[2] *Apache Jena Fuseki*. (n.d.). https://jena.apache.org/documentation/fuseki2/

[3] Berners-Lee, T. (2009, June). *Putting Government Data online - Design Issues*. https://www.w3.org/DesignIssues/GovData.html

[4] 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

[5] 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

[6] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., Yergeau, F., & W3C Recommendation. (2008, November 26). *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. https://www.w3.org/TR/xml/

[7] Brickley, D., & GUHA, R. V. (2014, February 25). *RDF Schema 1.1*. https://www.w3.org/TR/rdf11-schema/

[8] Brickley, D., & Miller, L. (2004, May 1). *FOAF Vocabulary Specification*. http://xmlns.com/foaf/0.1/

[9] Clarin Eric. (n.d.-a). *ParlaMint: Comparable Parliamentary Corpora*. https://github.com/clarin-eric/ParlaMint

[10] Clarin Eric. (n.d.-b). *ParlaMint: Towards comparable parliamentary corpora*. https://www.clarin.eu/parlamint

[11] *DCMI Metadata Terms*. (n.d.). https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[12] 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

[13] Dritsa, K. (n.d.). *Greek_Parliament_Proceedings_Dataset: Crawler and parser of the Greek Parliament proceedings (1989-2020)*. https://github.com/Dritsa-Konstantina/Greek_Parliament_Proceedings_Dataset

[14] Harris, S., Seaborne, A., & W3C Recommendation. (2013, March 21). *SPARQL 1.1 Query language*. https://www.w3.org/TR/sparql11-query/

[15] Hitzler, P., Krötzsch, M., & Rudolph, S. (2009). Foundations of Semantic Web Technologies. In *Chapman and Hall/CRC eBooks*. https://doi.org/10.1201/9781420090512

[16] *Lassila, O., & Swick, R. (1999). Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation*. (n.d.). https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[17] *Linked Data - design issues*. (n.d.). http://www.w3.org/DesignIssues/LinkedData.html

[18] *Linked Open Vocabularies (LOV)*. (n.d.). https://lov.linkeddata.es/dataset/lov

[19] *LinkedEP: Plenary debates of the European Parliament as Linked Open Data*. (n.d.). https://linkedpolitics.ops.few.vu.nl/web/html/home.html

[20] McCrae, J. P. (n.d.). *The linked Open Data cloud*. https://lod-cloud.net/

[21] Palmirani, M., Vitali, F., & OASIS Open. (n.d.). *OASIS LegalDocumentML (LegalDocML) TC*. https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=legaldocml

[22] Parr, T. (2013). *The Definitive ANTLR 4 reference*. Pragmatic Bookshelf. https://dl.acm.org/doi/10.5555/2501720

[23] RDFLib. (n.d.). *RDFLib is a Python library for working with RDF, a simple yet powerful language for representing information.* https://github.com/RDFLib/rdflib

[24] *Resource Description Framework (RDF)*. (2014, February 25). https://www.w3.org/RDF/

[25] TEI Consortium. (n.d.). *TEI: Text Encoding Initiative*. https://tei-c.org/

[26] *The debates of the Legislative Assembly | hansard.opennwt.ca*. (n.d.). https://hansard.opennwt.ca/debates/

[27] Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2016a). The debates of the European Parliament as Linked Open Data. *Semantic Web*, *8*(2), 271–281. https://doi.org/10.3233/sw-160227

[28] Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2016b). The debates of the European Parliament as Linked Open Data. *Semantic Web*, *8*(2), 271–281. https://doi.org/10.3233/sw-160227

[29] 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

[30] *xml.etree.ElementTree — The ElementTree XML API*. (n.d.). https://docs.python.org/3/library/xml.etree.elementtree.html

[31] Γενική Γραμματεία Νομικών και Κοινοβουλευτικών Θεμάτων. (n.d.). *Κυβερνήσεις από το 1909 έως σήμερα*. https://gslegal.gov.gr/?page_id=776&sort=time

[32] *Πρακτικό Συνεδρίασης Ολομέλειας του Ελληνικού Κοινοβουλίου - Παρασκευή 8 Ιουνίου 2018*. (2018, June 8). https://www.hellenicparliament.gr/UserFiles/a08fc2dd-61a9-4a83-b09a-09f4c564609d/es20180608_1.pdf

[33] *Συνεδριάσεις Ολομέλειας - Βουλή των Ελλήνων*. (n.d.). https://www.hellenicparliament.gr/Praktika/Synedriaseis-Olomeleias