



ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ  
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ  
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ  
«Eurostat's Finder»



Του φοιτητή  
Καμπουρίδη Θεόδωρου  
Αρ. Μητρώου: 2019060

Επιβλέπων  
Χαράλαμπος Μπράτσας  
Επίκουρος Καθηγητής

Ημερομηνία 12-09-2026

Τίτλος Δ.Ε. Eurostat's Finder

Κωδικός Δ.Ε. 23336

Ονοματεπώνυμο φοιτητή Καμπουρίδης Θεόδωρος-Χαράλαμπος

Ονοματεπώνυμο εισηγητή Μπράτσας Χαράλαμπος

Ημερομηνία ανάληψης Δ.Ε. 13-03-2025

Ημερομηνία περάτωσης Δ.Ε. 12-09-2026

*Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.*

*Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Καμπουρίδη Θεόδωρου-Χαράλαμπου που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης, ο συγγραφέας/δημιουργός εκχωρεί στο Διεθνές Πανεπιστήμιο της Ελλάδος άδεια χρήσης του δικαιώματος αναπαραγωγής, δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσης της εργασίας διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος. Η ανοικτή πρόσβαση στο πλήρες κείμενο της εργασίας, δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού, ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, πώληση, εμπορική χρήση, διανομή, έκδοση, μεταφόρτωση (downloading), ανάρτηση (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού.*

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

*Η εργασία αυτή είναι αφιερωμένη στην οικογένειά μου.*



## Πρόλογος

Η επιλογή της παρούσας διπλωματικής έγινε με γνώμονα τόσο το ενδιαφέρον μου για τα ανοιχτά δεδομένα (Open Data) και την ανάπτυξη διαδικτυακών εφαρμογών όσο και για τη συνεχώς αυξανόμενη χρήση στατιστικών δεδομένων για εξαγωγή συμπερασμάτων, συγκρίσεων και λήψη αποφάσεων.

Η Eurostat με τα πολλαπλά Application Programming Interfaces (APIs) της για παροχή δεδομένων και μεταδεδομένων μου έδωσε την ευκαιρία να περιηγηθώ για πρώτη φορά σε σύγχρονους, ανανεωμένους καταλόγους δεδομένων καθώς και να κατανοήσω τη διαδικασία ανάκτησης, διαχείρισης και ευρετηρίασης δεδομένων και των μεταδεδομένων τους, μέσω των APIs της, για ολοκληρωτική παροχή σημασιολογικών αποτελεσμάτων αντίστοιχων με τις απαιτήσεις του εκάστοτε χρήστη.

Η παραπάνω διαδικασία κατανόησης υπήρξε χρονοβόρα με διαδοχικές αναθεωρήσεις και συνεχείς βελτιώσεις για να επιτύχω το παρόν τελικό αποτέλεσμα. Ωστόσο, μου προσέφερε σημαντική εμπειρία στην κατανόηση και διαχείριση των μεταδεδομένων και την αξιοποίηση οργανωμένων δομών δεδομένων καθώς και στη μετέπειτα σημασιολογική αναπαράσταση και αξιοποίηση της ανακτημένης πληροφορίας μέσω τεχνολογιών Resource Description Framework (RDF).

## Περίληψη

Η Eurostat παρέχει ένα τεράστιο πλήθος στατιστικών δεδομένων και μεταδεδομένων της Ευρώπης καλύπτοντας τεράστιο όγκο θεμάτων τα οποία είναι προσβάσιμα μέσω των διαφορετικών APIs της. Το πλήθος των datasets σε συνδυασμό με τον απλοϊκό τρόπο αναζήτησής τους -βάσει μόνο του τίτλου- όπως και η αξιοποίηση των μεταδεδομένων εξακολουθούν να αποτελούν προκλήσεις για τον τελικό χρήστη. Σκοπός της παρούσας εργασίας είναι η ανάπτυξη της εφαρμογής Eurostat's Finder, με στόχο την αποδοτικότερη αναζήτηση συνόλων δεδομένων (datasets) και ανάκτηση δεδομένων, δημιουργία βάσης δεδομένων για αποδοτικότερη ευρετηρίαση, καθώς και χρήση μεταδεδομένων για τη σημασιολογική αναπαράσταση της ανακτημένης πληροφορίας σε μορφή RDF.

Για την επίτευξη του παραπάνω σκοπού αναπτύχθηκε ένας υβριδικός μηχανισμός εξαγωγής δεδομένων και μεταδεδομένων από τα Eurostat APIs αξιοποιώντας και την δυνατότητα της απευθείας αποθήκευσης σε αρχείο όπως και την άμεση αναπαράστασή τους στο web. Αναλυτικότερα, δημιουργήθηκε ένα αρχείο ευρετηρίασης με δύο πίνακες, ο πρώτος περιλαμβάνει για κάθε dataset πληροφορίες όπως τίτλος, θέμα, γονέας κόμβος, διαστάσεις, σημασιολογικές ετικέτες και αναγνωριστικό ID των μεταδεδομένων ενώ ο δεύτερος περιλαμβάνει το αναγνωριστικό ID με την περιγραφή (description) και τις έννοιες (concepts) του κάθε αρχείου μεταδεδομένων. Ο σημασιολογικός εμπλουτισμός πραγματοποιήθηκε μέσω της χρήσης όρων του EuroVoc, οι οποίοι αντιστοιχίστηκαν στα διαθέσιμα μεταδεδομένα. Η αρχιτεκτονική αυτή επιτρέπει την ταχύτατη αναζήτηση σε πολλαπλά πεδία, και όχι αποκλειστικά στον τίτλο, με τη χρήση του υβριδικού αλγορίθμου αξιολόγησης BM25.

Επιπλέον, τα επιλεγμένα στατιστικά αποτελέσματα αναπαρίστανται σημασιολογικά σε πραγματικό χρόνο (on-the-fly), εκμεταλλεύοντας το επίσημο λεξιλόγιο RDF Data Cube του Σημασιολογικού Ιστού ενώ παράλληλα παρέχεται η δυνατότητα λήψης τους από τους χρήστες. Μέσω αυτής της προσέγγισης αυξάνεται η σημασιολογική διαλειτουργικότητα της πληροφορίας που παρέχεται στον χρήστη αλλά εξαλείφεται η ανάγκη για τεράστιους τοπικούς αποθηκευτικούς χώρους.

# Eurostat’s Finder

Kampouridis Theodoros-Charalampos

## Abstract

Eurostat provides a vast amount of statistical data and metadata for Europe, covering a wide range of topics that are accessible through its various APIs. The sheer number of datasets, combined with the simplistic search method—based solely on the title—as well as the utilization of metadata continue to pose challenges for the end user. The purpose of this work is to develop the Eurostat’s Finder application with the aim of more efficient dataset search and data retrieval, using local indexing and metadata, as well as for the semantic representation of the retrieved information in RDF format.

To achieve this goal, a hybrid mechanism for extracting data and metadata from the Eurostat APIs was developed, leveraging both the ability to store data directly to a file and to represent it immediately on the web. More specifically, a local indexing file was created with two tables: the first includes information for each dataset, such as title, subject, parental node, dimensions, semantic labels and metadata ID while the second includes the ID along with the description and concepts for each dataset. Semantic enrichment was performed using terms from EuroVoc, which were automatically mapped to the available text. This approach allows advanced multi-field searching, and not exclusively by the title, by utilizing a hybrid BM25 ranking algorithm.

Furthermore, the retrieved statistical results are represented semantically on-the-fly using RDF technologies, adhering to the RDF Data Cube vocabulary of the Semantic Web. In addition, the capability of downloading these linked open data results is provided to users. This approach enhances the interoperability and machine readability of the provided information, while eliminating the need for massive local storage infrastructures.

## **Ευχαριστίες**

Θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου.

# Περιεχόμενα

|                                                                   |      |
|-------------------------------------------------------------------|------|
| Πρόλογος .....                                                    | v    |
| Περίληψη .....                                                    | vi   |
| Abstract.....                                                     | vii  |
| Ευχαριστίες .....                                                 | viii |
| Περιεχόμενα.....                                                  | ix   |
| Κατάλογος Σχημάτων.....                                           | xii  |
| Κατάλογος Πινάκων .....                                           | xii  |
| Συντομογραφίες.....                                               | xiv  |
| Κεφάλαιο 1ο: Εισαγωγή .....                                       | 1    |
| 1.1 Αντικείμενο Διπλωματικής .....                                | 1    |
| 1.2 Κίνητρο και Προβλήματα .....                                  | 1    |
| 1.3 Σκοπός και Στόχοι .....                                       | 2    |
| 1.4 Δομή Τεύχους .....                                            | 2    |
| Κεφάλαιο 2ο: Θεωρητικό Υπόβαθρο και Σχετικές Τεχνολογίες.....     | 4    |
| 2.1 Εισαγωγή.....                                                 | 4    |
| 2.1.1 Open Data και Eurostat .....                                | 4    |
| 2.1.2 Πρότυπο SDMX .....                                          | 4    |
| 2.1.3 ToC και ESMS Metadata της Eurostat.....                     | 4    |
| 2.1.4 Διεπαφές Προγραμματισμού Εφαρμογών (APIs) της Eurostat..... | 7    |
| 2.2 Σημασιολογικός Ιστός .....                                    | 7    |
| 2.2.1 Μοντέλο Δεδομένων RDF και Γλώσσα SPARQL .....               | 7    |
| 2.2.2 Πρότυπο SKOS και EuroVoc.....                               | 8    |
| 2.2.3 Λεξιλόγιο RDF Data Cube.....                                | 8    |
| 2.2.4 Λεξιλόγια Περιγραφής DCAT και DCTerms .....                 | 9    |
| 2.3 Ανάκτηση Πληροφορίας .....                                    | 9    |
| 2.3.1 Δημιουργία Ευρετηρίων σε Σχεσιακή Βάση.....                 | 10   |
| 2.3.2 Μηχανισμοί Βαθμολόγησης και Κατάταξης.....                  | 10   |
| 2.4 Σχετικές Εργασίες και Σύγχρονες Προσεγγίσεις .....            | 11   |
| 2.4.1 Προκλήσεις των Πυλών Ανοιχτών Κυβερνητικών Δεδομένων .....  | 11   |
| 2.4.2 Σημασιολογική Αναζήτηση σε Ανοιχτά Δεδομένα .....           | 11   |
| 2.4.3 Στατιστικά Δεδομένα και Λεξιλόγιο RDF Data Cube .....       | 12   |
| 2.4.4 Γενική Αποτίμηση της Εφαρμογής.....                         | 12   |

|              |                                                                                |    |
|--------------|--------------------------------------------------------------------------------|----|
| Κεφάλαιο 3ο: | Αρχιτεκτονική Συστήματος .....                                                 | 14 |
| 3.1          | Γενική Επισκόπηση .....                                                        | 14 |
| 3.2          | Μηχανισμός Συλλογής και Προεπεξεργασίας .....                                  | 15 |
| 3.2.1        | Ανάκτηση Καταλόγου και Μεταδεδομένων .....                                     | 15 |
| 3.2.2        | Εξόρυξη Δομής Δεδομένων .....                                                  | 15 |
| 3.3          | Έξυπνο Φιλτράρισμα και Μείωση Θορύβου .....                                    | 16 |
| 3.4          | Σημασιολογικός Εμπλουτισμός .....                                              | 16 |
| 3.5          | Σχεδιασμός Ενδιάμεσου Τοπικού Ευρετηρίου και Εισαγωγή στη Βάση Δεδομένων ..... | 16 |
| 3.6          | Σχεδιαστική Μοντελοποίηση Συστήματος .....                                     | 17 |
| 3.6.1        | Μοντέλο Οντοτήτων-Σχέσεων .....                                                | 17 |
| 3.6.2        | Διάγραμμα Αλληλουχίας Αναζήτησης .....                                         | 18 |
| 3.6.3        | Διάγραμμα Ενεργειών και Δυνατοτήτων Χρήστη .....                               | 18 |
| Κεφάλαιο 4ο: | Υλοποίηση Συστήματος .....                                                     | 20 |
| 4.1          | Τεχνολογίες και Εργαλεία Υλοποίησης .....                                      | 20 |
| 4.1.1        | Τεχνολογίες Υποβάθρου και Διαχείριση Δεδομένων .....                           | 20 |
| 4.1.2        | Τεχνολογίες Διεπαφής Χρήστη .....                                              | 20 |
| 4.1.3        | Εξωτερικά Πρότυπα .....                                                        | 20 |
| 4.2          | Κατασκευή και Δόμηση των Πινάκων .....                                         | 21 |
| 4.3          | Μηχανή Αναζήτησης .....                                                        | 24 |
| 4.4          | Διεπαφή Χρήστη .....                                                           | 26 |
| 4.4.1        | Δυναμικό Φιλτράρισμα .....                                                     | 26 |
| 4.4.2        | Οπτικοποίηση Δεδομένων σε Πίνακα .....                                         | 27 |
| 4.5          | Παραγωγή Σημασιολογικής Δομής .....                                            | 28 |
| 4.6          | Μηχανισμός Διαμεσολάβησης και Διαχείρισης CORS .....                           | 33 |
| Κεφάλαιο 5ο: | Αποτελέσματα και Αξιολόγηση .....                                              | 34 |
| 5.1          | Σενάρια Χρήσης .....                                                           | 34 |
| 5.1.1        | Σενάριο Άμεσης Λεκτικής Ταύτισης .....                                         | 34 |
| 5.1.2        | Σενάριο Ταύτισης μέσω Σημασιολογικής Αναζήτησης .....                          | 37 |
| 5.1.3        | Ποσοτική Αξιολόγηση Αποτελεσμάτων .....                                        | 42 |
| 5.2          | Σύγκριση με το Επίσημο Εργαλείο της Eurostat .....                             | 43 |
| 5.2.1        | Σχεδιασμός Διεπαφής Χρήστη .....                                               | 43 |
| 5.2.2        | Διαφορές στην Μηχανή Αναζήτησης .....                                          | 44 |
| 5.2.3        | Λήψη Στατιστικών Δεδομένων .....                                               | 45 |
| 5.3          | Αξιολόγηση Απόδοσης .....                                                      | 45 |
| 5.3.1        | Βελτιστοποίηση Αποθηκευτικού Χώρου .....                                       | 45 |

|                    |                                                                                |    |
|--------------------|--------------------------------------------------------------------------------|----|
| 5.3.2              | Ταχύτητα Ανάκτησης και Υβριδική Αναζήτηση.....                                 | 46 |
| 5.3.3              | Δυναμική Παραγωγή Σημασιολογικών Δεδομένων .....                               | 46 |
| 5.4                | Ποσοτική Αποτίμηση και Μετρικές Συστήματος .....                               | 47 |
| Κεφάλαιο 6ο:       | Συμπεράσματα και Προτάσεις Βελτίωσης.....                                      | 48 |
| 6.1                | Σύνοψη και Συμπεράσματα.....                                                   | 48 |
| 6.1.1              | Περιορισμοί Υφιστάμενης Υλοποίησης .....                                       | 48 |
| 6.2                | Μελλοντικές Επεκτάσεις.....                                                    | 49 |
| 6.2.1              | Ενσωμάτωση Μηχανισμών Επεξεργασίας Φυσικής Γλώσσας .....                       | 49 |
| 6.2.2              | Δυναμικός Αλγόριθμος Εντοπισμού Μετρήσιμου Μεγέθους.....                       | 49 |
| 6.2.3              | Παροχή Δημόσιου REST API .....                                                 | 49 |
| 6.2.4              | Υποστήριξη Άλλων Γλωσσών.....                                                  | 50 |
| 6.2.5              | Αυτοματοποιημένος Μηχανισμός Συγχρονισμού Δεδομένων.....                       | 50 |
| BIBΛΙΟΓΡΑΦΙΑ ..... |                                                                                | 51 |
| ΠΑΡΑΡΤΗΜΑ Α :      | Πλήρες Παράδειγμα Παραγόμενου Αρχείου Σημασιολογικού Κύβου RDF Data Cube ..... | 53 |
| ΠΑΡΑΡΤΗΜΑ Β :      | Παραγόμενη Δομή DSD.....                                                       | 66 |

## Κατάλογος Σχημάτων

|                                                                                                      |    |
|------------------------------------------------------------------------------------------------------|----|
| Σχήμα 2.1: Εμφάνιση του αρχικού μέρους του καταλόγου ToC στον ιστότοπο της Eurostat .....            | 5  |
| Σχήμα 2.2: Αρχή ιεραρχικού καταλόγου ToC σε μορφή xml.....                                           | 6  |
| Σχήμα 2.3: Ενότητα 3.1 του ESMS .....                                                                | 6  |
| Σχήμα 2.4: Ενότητα 3.4 του ESMS .....                                                                | 7  |
| Σχήμα 2.5: Βασική δομή λεξιλογίου RDF Data Cube [12] .....                                           | 9  |
| Σχήμα 3.1: Διάγραμμα Αρχιτεκτονικής Συστήματος.....                                                  | 14 |
| Σχήμα 3.2: Διάγραμμα Οντοτήτων-Συσχετίσεων.....                                                      | 17 |
| Σχήμα 3.3: Διάγραμμα Αλληλουχίας Αναζήτησης.....                                                     | 18 |
| Σχήμα 3.4: Διάγραμμα Ενεργειών Χρήστη .....                                                          | 19 |
| Σχήμα 4.1: Τεχνολογική Στοιβά Υλοποίησης της Eurostat's Finder.....                                  | 21 |
| Σχήμα 4.2: Εξαγωγή βασικών αναγνωριστικών από ToC .....                                              | 21 |
| Σχήμα 4.3: Συνάρτηση ανάκτησης description και concepts.....                                         | 22 |
| Σχήμα 4.4: Συνένωση όλων των πεδίων στην μεταβλητή fullText .....                                    | 23 |
| Σχήμα 4.5: Συνάρτηση ταύτισης και εξαγωγής όρων του EuroVoc .....                                    | 23 |
| Σχήμα 4.6: Δημιουργία πινάκων στην mysql-converter.php .....                                         | 24 |
| Σχήμα 4.7: Υλοποίηση αλγορίθμου κατάταξης και ενίσχυσης πεδίων μέσω σύνθετου SQL ερωτήματος .....    | 25 |
| Σχήμα 4.8: Αρχική Διεπαφή Χρήστη .....                                                               | 26 |
| Σχήμα 4.9: Τμήμα του UI που δείχνει την εμφάνιση των φίλτρων .....                                   | 27 |
| Σχήμα 4.10: Οπτικοποίηση δεδομένων σε πίνακα βάσει φίλτρων .....                                     | 28 |
| Σχήμα 4.11: DSD δομή σε μορφή .ttl.....                                                              | 29 |
| Σχήμα 4.12: Δήλωση prefixes, τίτλου και περιγραφής σε παραγόμενο αρχείο RDF με παρατηρήσεις .....    | 30 |
| Σχήμα 4.13: Δήλωση ετικετών EuroVoc και αρχή των παρατηρήσεων σε αρχείο RDF με παρατηρήσεις .....    | 31 |
| Σχήμα 4.14: Δήλωση Codelabels σε παραγόμενο RDF με παρατηρήσεις .....                                | 31 |
| Σχήμα 4.15: Διάγραμμα RDF Data Cube ενός τυχαίου παραδείγματος σε τυχαίο dataset .....               | 32 |
| Σχήμα 5.1: Σενάριο άμεσης λεκτικής ταύτισης με τίτλο .....                                           | 34 |
| Σχήμα 5.2: Εμφάνιση score του παραπάνω παραδείγματος .....                                           | 35 |
| Σχήμα 5.3: Σενάριο εισαγωγής αναγνωριστικού ID .....                                                 | 36 |
| Σχήμα 5.4: Εμφάνιση score του παραπάνω παραδείγματος .....                                           | 36 |
| Σχήμα 5.5: Σενάριο εισαγωγής όρου εκτός τίτλου .....                                                 | 37 |
| Σχήμα 5.6: Εμφάνιση score του παραπάνω παραδείγματος .....                                           | 38 |
| Σχήμα 5.7: Σενάριο εισαγωγής δύο λέξεων εκτός τίτλου στην μηχανή αναζήτησης .....                    | 39 |
| Σχήμα 5.8: Εμφάνιση score του παραπάνω παραδείγματος .....                                           | 39 |
| Σχήμα 5.9: Αρχείο ESMS sbs_pu .....                                                                  | 40 |
| Σχήμα 5.10: Εισαγωγή του όρου «food» .....                                                           | 40 |
| Σχήμα 5.11: Εισαγωγή του όρου «additives» .....                                                      | 41 |
| Σχήμα 5.12: Εισαγωγή των όρων «food additives».....                                                  | 41 |
| Σχήμα 5.13: Εισαγωγή όρου «pregnancy» στην μηχανή αναζήτησης επίσημου εργαλείου της Eurostat .....   | 44 |
| Σχήμα 5.14: Εισαγωγή όρων «coal market» στην μηχανή αναζήτησης επίσημου εργαλείου της Eurostat ..... | 45 |

## Κατάλογος Πινάκων

|                                                                                                     |    |
|-----------------------------------------------------------------------------------------------------|----|
| Πίνακας 2.1: Συγκριτική Αξιολόγηση Χαρακτηριστικών Συστήματος και Υφιστάμενων Προσεγγίσεων .....    | 13 |
| Πίνακας 4.1: Αντιστοίχιση των πεδίων της Eurostat με επίσημους όρους σημασιολογικής ιδιότητας ..... | 32 |
| Πίνακας 5.1: Σύγκριση αποτελεσματικότητας απλής και υβριδικής BM25 αναζήτησης .....                 | 42 |
| Πίνακας 5.2: Χρόνοι απόκρισης βασικών λειτουργιών της Eurostat's Finder .....                       | 47 |

## Συντομογραφίες

|       |                                        |
|-------|----------------------------------------|
| Δ.Ε.  | Διπλωματική Εργασία                    |
| ΔΙΠΑΕ | Διεθνές Πανεπιστήμιο Ελλάδος           |
| API   | Application Programming Interface      |
| RDF   | Resource Description Framework         |
| SDMX  | Statistical Data and Metadata eXchange |
| ΕΕ    | Ευρωπαϊκή Ένωση                        |
| ToC   | Table of Contents                      |
| ESMS  | Euro SDMX Metadata Structure           |
| W3C   | World Wide Web Consortium              |
| SKOS  | Simple Knowledge Organization System   |
| UI    | User Interface                         |
| DOM   | Document Object Model                  |
| AJAX  | Asynchronous JavaScript and XML        |
| RAM   | Random Access Memory                   |
| NLP   | Natural Language Processing            |
| FK    | Foreign Key                            |
| CORS  | Cross-Origin Resource Sharing          |
| OGD   | Open Government Data                   |

# Κεφάλαιο 1ο: Εισαγωγή

## 1.1 Αντικείμενο Διπλωματικής

Στην σύγχρονη ψηφιακή εποχή, η διάθεση και η χρήση δεδομένων αυξάνονται ολοένα και περισσότερο, με τα Ανοιχτά Δεδομένα (Open Data) να χρησιμοποιούνται με εκθετικούς ρυθμούς. Διεθνείς οργανισμοί αξιοποιώντας την τάση αυτήν, δημοσιεύουν τεράστιους όγκους στατιστικών δεδομένων, ενισχύοντας την διαφάνεια της έρευνας [16]. Συγκεκριμένα για την παρούσα διπλωματική, η Eurostat αποτελεί τον βασικό ευρωπαϊκό φορέα παροχής στατιστικών πληροφοριών για τα κράτη της Ευρωπαϊκής Ένωσης, με βασικότερο χαρακτηριστικό την αξιοπιστία των παρεχόμενων στοιχείων της.

Ωστόσο, ένας «μύθος» που ακολουθεί τα Ανοιχτά Δεδομένα, είναι η ταύτιση της δημόσιας παροχής δεδομένων με το σκεπτικό της εύκολης ανάκτησης και αξιοποίησης από τον καθένα. Ο δομημένος σκελετός των δεδομένων, ο τεράστιος όγκος τους και η χρήση εξειδικευμένων όρων δυσχεραίνει τις διαδικασίες αυτές για το ευρύ κοινό [15]. Συνεπώς, μετά την αναγνώριση του προβλήματος, η παρούσα εργασία αποσκοπεί στη μείωση και αντιμετώπιση του προβλήματος. Αντικείμενο της διπλωματικής είναι ο αρχικός σχεδιασμός και η μετέπειτα ανάπτυξη ενός πρωτότυπου συστήματος ανάκτησης και σημασιολογικού εμπλουτισμού δεδομένων φιλικό προς τον μέσο χρήστη, με όνομα «Eurostat's Finder».

Αναλυτικότερα, το σύστημα λειτουργεί ως «ενδιάμεσος» μεταξύ των στοιχείων που παρέχει η Eurostat μέσω των APIs της και του τελικού χρήστη μέσω της διεπαφής χρήστη, με βασικό πλάνο την ενισχυμένη σημασιολογικά αναζήτηση των ανοικτών συνόλων δεδομένων (datasets) αλλά και την δυναμική μετατροπή τους σε πρότυπα του Σημασιολογικού Ιστού (Linked Open Data) [18].

## 1.2 Κίνητρο και Προβλήματα

Το κίνητρο της εργασίας εντοπίζεται στον προγενέστερο εντοπισμό των αδυναμιών στο -ήδη υπάρχον- επίσημο εργαλείο της Eurostat αλλά και γενικά σε παρόμοια συστήματα αναζήτησης, και έπειτα από τους περιορισμούς φιλοξενίας του ακαδημαϊκού εξυπηρετητή. Συγκεκριμένα, τα προβλήματα που φανερώθηκαν και συνέβαλαν ως αφορμή ανάπτυξης μίας εφαρμογής σαν το Eurostat's Finder είναι τα εξής:

- **Απλή και περιορισμένη αναζήτηση:** Στην πλατφόρμα της Eurostat (Data Browser), η αναζήτηση ταυτίζεται απόλυτα είτε με λεκτική ταύτιση στον τίτλο του συνόλου δεδομένων ή στην ταύτιση με το μοναδικό αλφαριθμητικό αναγνωριστικό του (ID). Συνεπώς, δεν δίνεται η δυνατότητα πλοήγησης με σχετικές, ευρύτερες έννοιες/λέξεις-κλειδιά παρά μόνον αν περιέχονται στον τίτλο, καταλήγοντας συχνά σε μηδενικά αποτελέσματα και σε χειροκίνητη πλοήγηση στον πίνακα των περιεχομένων (Table of Contents), που περιέχει όλα τα datasets, δημιουργώντας εν τέλει συχνές υπερφορτώσεις πληροφορίας [17].
- **Απουσία σημασιολογικής δομής:** Ύστερα από τον επιτυχή εντοπισμό των δεδομένων, η εξαγωγή τους από το επίσημο εργαλείο της Eurostat πραγματοποιείται μέσω συμβατών μορφώσεων, όπως CSV, Excel, JSON, που είναι κατάλληλα για ανθρώπινη ανάγνωση και κατανόηση. Όμως στερούνται ρητής σημασιολογικής αναπαράστασης και διασύνδεσης. Παρότι μορφές όπως JSON, CSV και XML είναι αναγνώσιμες από μηχανές (machine-readability), δεν προσφέρουν την άμεση σημασιολογική διαλειτουργικότητα που απαιτεί ο Σημασιολογικός Ιστός.
- **Τεχνικοί περιορισμοί φιλοξενίας:** Η δημιουργία μιας εμπλουτισμένης -σημασιολογικά- βάσης δεδομένων με στατιστικά από την Eurostat, εγείρει απαιτήσεις γύρω από τον αποθηκευτικό χώρο (disk quotas) και από την μεγάλη επεξεργαστική ισχύ (RAM) που θα χρειαστούν. Η ανάπτυξή του σε ακαδημαϊκό περιβάλλον φιλοξενίας, με αυστηρούς περιορισμούς στις παραπάνω παραμέτρους αποτέλεσε μια μείζονα αρχιτεκτονική πρόκληση.

### 1.3 Σκοπός και Στόχοι

Σκοπός της διπλωματικής εργασίας είναι η υπέρβαση των παραπάνω περιορισμών και προβλημάτων, μέσω σύγχρονων τεχνολογιών, δημιουργώντας μία ολοκληρωμένη και άκρως αποδοτική διαδικτυακή εφαρμογή.

Για να επιτευχθεί αυτός ο σκοπός, υλοποιήθηκαν οι ακόλουθοι στόχοι:

- **Ανάπτυξη αυτοματοποιημένου μηχανισμού εξόρυξης πληροφοριών:** Η ύπαρξη ασύγχρονου μηχανισμού για την συλλογή των πληροφοριών των datasets κρίνεται επιτακτική. Το data builder αρχείο υλοποιείται σε περιβάλλον Node.js, αναλαμβάνει να ανακτά αυτομάτως τα δεδομένα και μεταδεδωμένα των datasets ενώ σε περιπτώσεις, όπως τα μεταδεδωμένα ESMS, εφαρμόζει φιλτράρισμα πληροφορίας με σκοπό την εξάλειψη στατιστικού «θορύβου».
- **Σημασιολογικός εμπλουτισμός:** Παράλληλα με την διάρκεια της αυτοματοποιημένης εξόρυξης των πληροφοριών, κρίνεται σημαντικό ο σημασιολογικός εμπλουτισμός των datasets βάσει του επίσημου ευρωπαϊκού λεξιλογίου EuroVoc, προκειμένου να αποδοθεί περαιτέρω θεματικό βάθος μέσω των ετικετών.
- **Αποδοτική διαχείριση αποθήκευσης βάσης δεδομένων και βελτιστοποίηση αναζήτησης:** Αποδοτικός αρχιτεκτονικός σχεδιασμός της κανονικοποιημένης βάσης δεδομένων ώστε να ξεπεραστούν τα τεχνικά θέματα των περιορισμών του διακομιστή, κυρίως σε αποθηκευτικό χώρο. Επιπρόσθετα, υλοποίηση και εφαρμογή υβριδικού αλγορίθμου κατανομής διαφορετικών βαρών στα πεδία της βάσης δεδομένων με σκοπό την αποδοτική κατανομή λεκτικών ταυτίσεων βάσει σημαντικότητας. Συνεπώς, βελτιώνοντας την σχετικότητα μεταξύ πιθανών αποτελεσμάτων και τους εισαγόμενους όρους των χρηστών.
- **Δυναμική παραγωγή αποτελεσμάτων ανοιχτών δεδομένων:** Η υλοποίηση συστήματος δυναμικής μετατροπής των στατιστικών δεδομένων σε πραγματικό χρόνο, βάσει των φίλτρων που επέλεξε ο εκάστοτε χρήστης, στην επίσημη μορφή του W3C (World Wide Web Consortium) RDF Data Cube Vocabulary, ελαχιστοποιώντας το αποτύπωμα αποθηκευτικού χώρου (zero storage footprint) και διασφαλίζοντας την διαλειτουργικότητα των μηχανών.

Με την επίτευξη των παραπάνω στόχων, η παρούσα εργασία επιχειρεί να γεφυρώσει το χάσμα μεταξύ των συμβατικών APIs και των τεχνολογιών του Σημασιολογικού Ιστού, προτείνοντας ένα καινοτόμο, ενδιάμεσο επίπεδο που προσφέρει σημασιολογική διαλειτουργικότητα και αναζήτηση βελτιωμένης ακρίβειας, ενώ παράλληλα η δυναμική παραγωγή των δεδομένων σε RDF, ξεπερνάει την ανάγκη μεγάλων υποδομών αποθήκευσης .

### 1.4 Δομή Τεύχους

Η παρούσα διπλωματική εργασία αναπτύσσεται σε έξι κεφάλαια, με την οργάνωση να έχει ως εξής:

- Στο Κεφάλαιο 1, παρουσιάζεται η γενική εικόνα, το αντικείμενο της εργασίας, οι περιορισμοί, τα κίνητρα, καθώς και οι τεχνολογικοί στόχοι που έχουν τεθεί.
- Στο Κεφάλαιο 2, περιγράφεται το θεωρητικό υπόβαθρο της έρευνας. Αναλύονται συγκεκριμένα πρότυπα που χρησιμοποιήθηκαν και μερικές κύριες αρχές ανάκτησης πληροφορίας.
- Στο Κεφάλαιο 3, προβάλλεται η αρχιτεκτονική της εφαρμογής, ξεκινώντας από τη ροή των δεδομένων και φτάνοντας έως την αντιμετώπιση των περιορισμών φιλοξενίας.
- Στο Κεφάλαιο 4, αναλύεται εκτενώς η προγραμματιστική υλοποίηση της αρχιτεκτονικής που εξηγήθηκε στο παραπάνω κεφάλαιο. Απεικονίζονται τμήματα κώδικα που συσχετίζονται με την κατασκευή της βάσης δεδομένων και την διεπαφή χρήστη, περιέχοντας τις παραγωγές της μηχανής αναζήτησης και του Σημασιολογικού Κύβου.
- Στο Κεφάλαιο 5, αξιολογείται η αποδοτικότητα της εφαρμογής βάσει σεναρίων χρήσης και διενεργούνται παράλληλες συγκρίσεις με το επίσημο εργαλείο της Eurostat.

- Στο Κεφάλαιο 6, συνοψίζονται τα τελικά συμπεράσματα και προτείνονται νέες ιδέες μελλοντικών επεκτάσεων της εφαρμογής.

## Κεφάλαιο 2ο: Θεωρητικό Υπόβαθρο και Σχετικές Τεχνολογίες

### 2.1 Εισαγωγή

#### 2.1.1 Open Data και Eurostat

Open Data ή Ανοιχτά Δεδομένα χαρακτηρίζονται τα δεδομένα τα οποία μπορούν να χρησιμοποιηθούν ελεύθερα δίχως έγκριση και είναι μόνιμα διαθέσιμα προς χρήση. Το βασικότερο προτέρημα τους είναι η διαλειτουργικότητα που προσφέρουν μεταξύ διαφορετικών συστημάτων και οργανισμών επιτυγχάνοντας την ολοκλήρωση πολύπλοκων και μεγάλων συστημάτων [1]. Σε Ευρωπαϊκό επίπεδο, με την οδηγία 2019/1024 της Ευρωπαϊκής Ένωσης (ΕΕ), επισφραγίστηκε η στρατηγική σημασία των Open Data προωθώντας την επαναχρησιμοποίηση πληροφοριών δημοσίου τομέα [2].

Ωστόσο σε πολλές περιπτώσεις, όπως και αυτή της Eurostat, η ελεύθερη διανομή των δεδομένων δεν συνεπάγεται την ευκολία ανάκτησής τους από χρήστες. Εξειδικευμένες μορφές αρχείων, ακανόνιστες δομές δεδομένων και απουσία εργαλείων αναζήτησης είναι τα πιο συνήθη προβλήματα που καλούνται να αντιμετωπίσουν οι χρήστες [3]. Βάσει αυτών των προβλημάτων γεννήθηκε η ιδέα ύπαρξης της ενδιάμεσης εφαρμογής Eurostat's Finder που θα λειτουργεί ως δίαυλος μεταξύ της πλατφόρμας της Eurostat και των χρηστών, ικανή να διαχειριστεί και να μετατρέψει τις ογκώδεις και δαιδαλώδεις δομές δεδομένων της Eurostat σε εύκολα προσβάσιμα δεδομένα στο ευρύ κοινό.

#### 2.1.2 Πρότυπο SDMX

Το SDMX, ακρωνύμιο του Statistical Data and Metadata eXchange, αποτελεί το διεθνές πρότυπο ISO 17369:2013 και χρησιμοποιείται για περιγραφή δεδομένων και μεταδεδομένων διευκολύνοντας την κοινή χρήση μεταξύ διαφορετικών οργανισμών [4] ενώ αποτελείται από τρεις βασικούς πυλώνες: τεχνικά πρότυπα, στατιστικές κατευθυντήριες γραμμές και αρχιτεκτονική και εργαλεία πληροφορικής [5]. Η μοντελοποίηση των στατιστικών πληροφοριών γίνεται με τυποποιημένο τρόπο και χρήση συγκεκριμένου γλωσσариού ώστε να επιτυγχάνεται αποδοτικά η κεντρική διαχείριση και η τυποποιημένη πρόσβαση στα δεδομένα ενώ οι διαστάσεις (dimensions) και οι μετρήσεις (measures) αναπαριστούν τις περιγραφές και τις τιμές που έχουν μετρηθεί αντίστοιχα [4].

Η πλήρης οργάνωση και ανάκτηση ενός συνόλου δεδομένων ακολουθεί μία αυστηρή, ιεραρχική ροή πληροφορίας. Αρχικά, με την προσθήκη του id στο κατάλληλο API call, εντοπίζεται η Ροή Δεδομένων (Dataflow), η οποία παραπέμπει στον μοναδικό Ορισμό Δομής Δεδομένων (DSD). Το DSD με την σειρά του εμπεριέχει όλες τις επιμέρους διαστάσεις (dimensions), οι οποίες συνδέονται η καθεμιά σε αυστηρά σε μία Λίστα Κωδικών (Codelist). Εκεί παρέχεται όλο το αποδεκτό σύνολο τιμών που μπορεί να λάβει η συγκεκριμένη διάσταση. Συνεπώς, η ιεραρχική αυτή ροή διασφαλίζει την τυποποίηση των μεταδιδόμενων στατιστικών στοιχείων [4].

#### 2.1.3 ToC και ESMS Metadata της Eurostat

Η πλατφόρμα της Eurostat διαθέτει μια οργανωμένη, ιεραρχική δομή δεδομένων η οποία ονομάζεται Table of Contents (ToC) και βασίζεται κυρίως στον συνεχή θεματικό διαχωρισμό των δεδομένων μέχρι τον εντοπισμό των επιμέρους datasets. Εκεί, βρίσκεται ένα πλήθος datasets με συγκεκριμένο τίτλο και

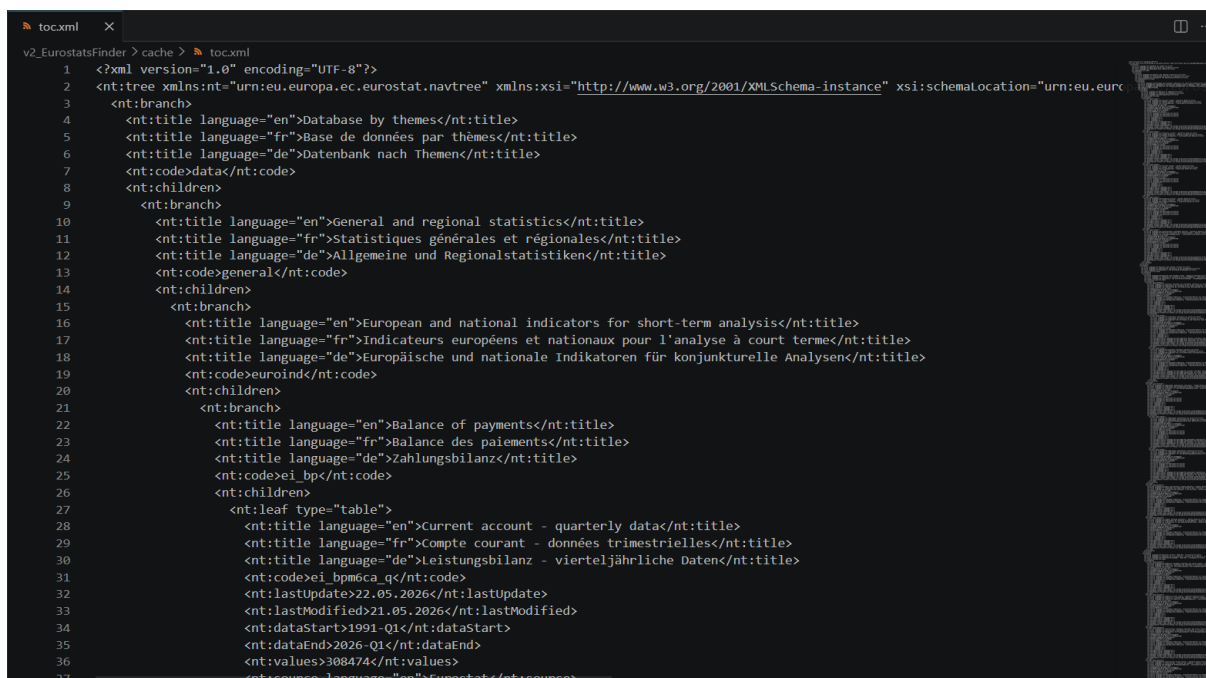
μοναδικό αλφαριθμητικό κωδικό id σε παρένθεση, περιέχοντας πληροφορίες για συγκεκριμένες μεταβλητές, όπως απεικονίζεται στο σχήμα 2.1.

## Database

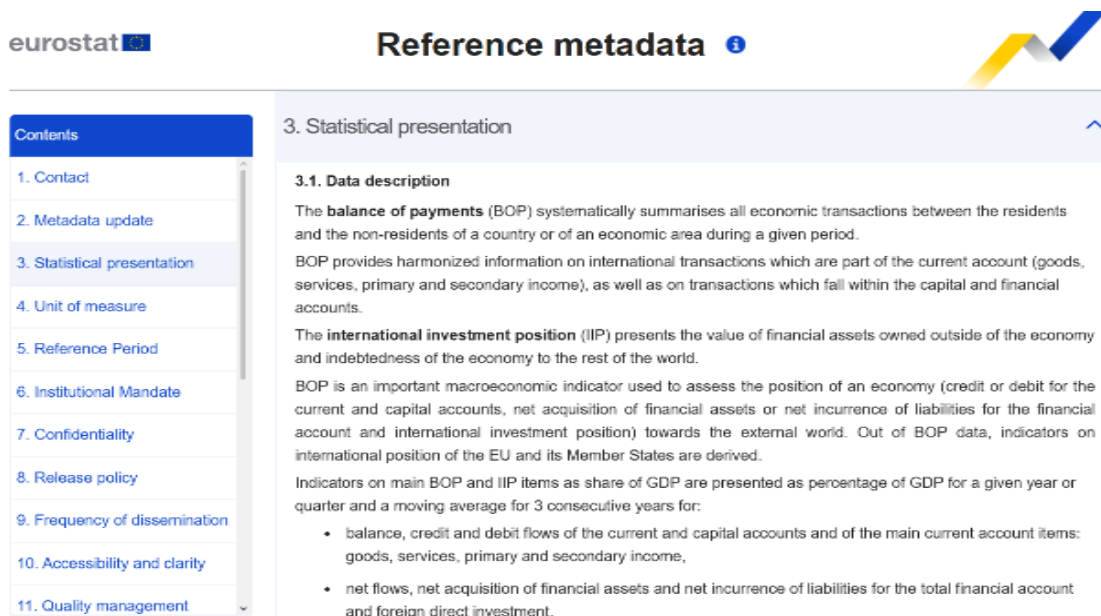


Σχήμα 2.1: Εμφάνιση του αρχικού μέρους του καταλόγου ToC στον ιστότοπο της Eurostat

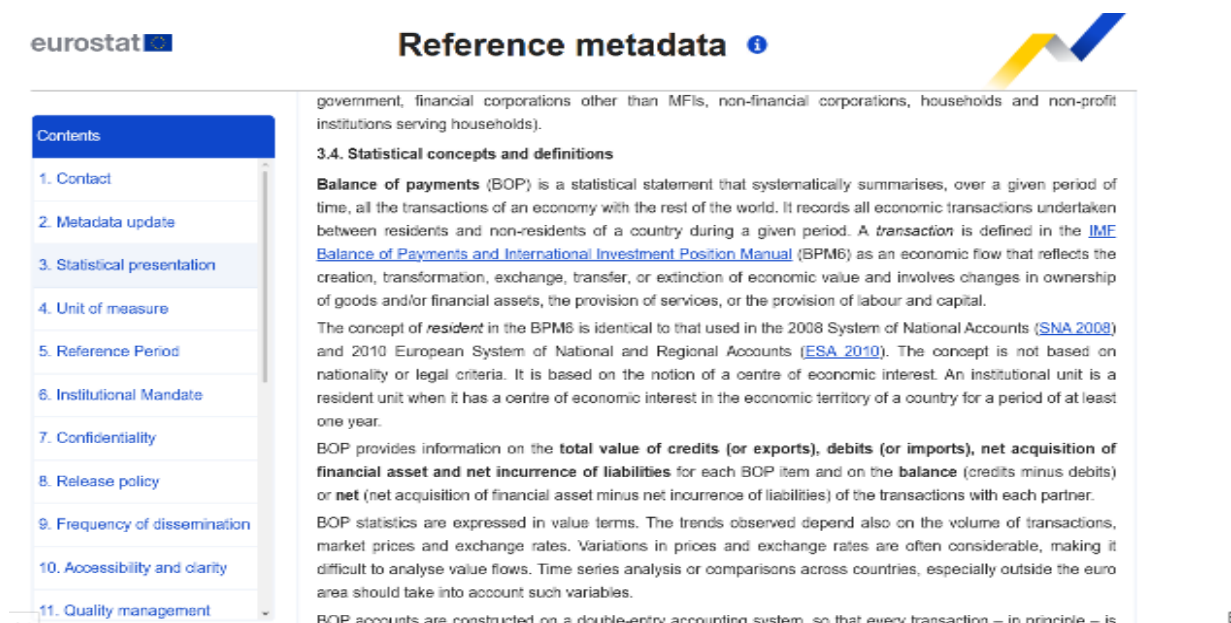
Στο Σχήμα 2.1, φαίνεται επίσης ένα εικονίδιο δίπλα από τα περισσότερα σύνολα δεδομένων υπάρχει ένα εικονίδιο με το γράμμα «M» ( από τον όρο Metadata) που παραπέμπει σε σελίδες html που ακολουθούν το πρότυπο ESMS (Euro SDMX Metadata Structure) και αναφέρουν όλα τα μεταδεδομένα του αντίστοιχου dataset ταξινομημένα σε θεματικές κατηγορίες [6]. Στην παρούσα εργασία, πραγματοποιείται αυτοματοποιημένη ανάκτηση του ToC σε μορφή xml, όπως απεικονίζεται στο σχήμα 2.2. Στην συνέχεια, γίνεται δυναμική εύρεση του ESMS συνδέσμου, εφόσον υπάρχει, και εξάγεται και αποθηκεύεται το περιεχόμενο των ενότητων 3.1 Data description και 3.4 Statistical concepts and definitions με σκοπό την ανάπτυξη σημασιολογικά εμπλουτισμένου ευρετηρίου αναζήτησης.



Σχήμα 2.2: Αρχή ιεραρχικού καταλόγου ToC σε μορφή xml



Σχήμα 2.3: Ενότητα 3.1 του ESMS



Σχήμα 2.4: Ενότητα 3.4 του ESMS

### 2.1.4 Διεπαφές Προγραμματισμού Εφαρμογών (APIs) της Eurostat

Το σύνολο της ακατέργαστης πληροφορίας που αξιοποιείται σε αυτήν την εφαρμογή ανακτάται αποκλειστικά από τις επίσημες διεπαφές της Eurostat. Στην παρούσα εφαρμογή αξιοποιούνται τρία από τα τέσσερα διακριτά APIs, με το καθένα να έχει διαφορετική χρήση:

- **Catalogue:** Χρησιμοποιείται στην εκτός-σύνδεσης προετοιμασία των δεδομένων για την λήψη του πλήρους καταλόγου ToC σε μορφή xml, από όπου εξάγονται βασικές πληροφορίες όπως ο τίτλος, τα θέματα, το αναγνωριστικό id και οι σύνδεσμοι των μεταδεδομένων. Επίσης, βρίσκει μία δεύτερη χρήση στην αναζήτηση dataset μέσω πλοήγησης στο δέντρο ToC - βάσει θεματικής- αντί της αναζήτησης λέξεων κλειδιών.
- **SDMX3.0:** Παρέχει όλες τις πληροφορίες σχετικά με τις αυστηρές δομημένες σχεσιακές πληροφορίες. Μέσω αυτού ανακτώνται τα Dataflows, τα DSDs, τα Dimensions και τα Codelists του εκάστοτε dataset.
- **Statistics1.0:** Χρησιμοποιείται από την διεπαφή του χρήστη για την ανάκτηση μόνο των ενεργών τιμών των φίλτρων, απορρίπτοντας τις τιμές που δεν εμφανίζουν δεδομένα. Δηλαδή «φιλτράρει» τα δεδομένα που λαμβάνονται από τα Codelists ώστε να εμφανίζονται μόνο τα έγκυρα για το εκάστοτε dataset.

## 2.2 Σημασιολογικός Ιστός

### 2.2.1 Μοντέλο Δεδομένων RDF και Γλώσσα SPARQL

Ο Σημασιολογικός Ιστός (Semantic Web) αποτελεί τον εξελικτικό στόχο του Παγκόσμιου Ιστού, αποσκοπώντας στην δημιουργία ενός δικτύου δεδομένων ικανών να αναγνωστούν και να επεξεργαστούν αυτομάτως από μηχανές, σε αντίθεση με τον παραδοσιακό ιστό που είναι σχεδιασμένος για την ανθρώπινη κατανόηση και επεξεργασία [7]. Ο συνδυασμός κρίκος αυτής της εξέλιξης είναι το RDF της Κοινοπραξίας του Παγκόσμιου Ιστού (World Wide Web Consortium – W3C), ένα κοινό τυποποιημένο μοντέλο αναπαράστασης που λειτουργεί ως «γραμματική» του Σημασιολογικού Ιστού [8]. Συγκεκριμένα, δομεί την πληροφορία με την χρήση απλών προτάσεων διαχωρίζοντάς τες σε

υποκείμενο (subject), κατηγορημα (predicate) και αντικείμενο (object), αναφερόμενες και ως τριάδες (triples) [8]. Για την ανάκτηση και αναζήτηση τέτοιων δεδομένων, η κοινότητα του W3C έχει καθιερώσει την γλώσσα ερωτημάτων SPARQL, με δυνατότητες υποβολής ερωτημάτων καθώς και συνδέσεις και διαζεύξεις μεταξύ τους [9].

Συνεπώς, η ενσωμάτωση του μοντέλου RDF αναβαθμίζει το αρχιτεκτονικό υπόβαθρο της εργασίας από ένα απλό επίπεδο διαχείρισης δεδομένων σε ένα άκρως σημασιολογικό. Αν και η γλώσσα SPARQL αποτελεί το πρότυπο ερωτημάτων στον Σημασιολογικό Ιστό, η παρούσα εφαρμογή εστιάζει στην άμεση παραγωγή και μοντελοποίηση των στατιστικών στοιχείων σε τριάδες (RDF Triples), προσδίδοντας εννοιολογική δομή και δυνατότητα μελλοντικής επαναχρησιμοποίησης στον ευρύτερο Σημασιολογικό Ιστό.

### 2.2.2 Πρότυπο SKOS και EuroVoc

Παρότι το RDF παρέχει τον βασικό σκελετό για την δημιουργία σημασιολογικών τριάδων, η W3C έχει θεσπίσει το εξειδικευμένο πρότυπο Simple Knowledge Organization System (SKOS) για την οργάνωση των εννοιών σε δομημένα λεξιλόγια και θησαυρούς [10]. Το SKOS αποτελεί ένα κοινό μοντέλο δεδομένων για κοινοποίηση και σύνδεση συστημάτων οργάνωσης γνώσης στον Σημασιολογικό Ιστό, με βασική δυνατότητα την συσχέτιση εννοιών μέσω της χρήσης προτιμώμενου (prefLabel) και εναλλακτικού όρου (altLabel). Επιπρόσθετα, το SKOS εξυπηρετεί την ιεραρχική δόμηση της γνώσης, επιτρέποντας την ταξινόμηση των δεδομένων σε λογικά, διασυνδεδεμένα δέντρα.

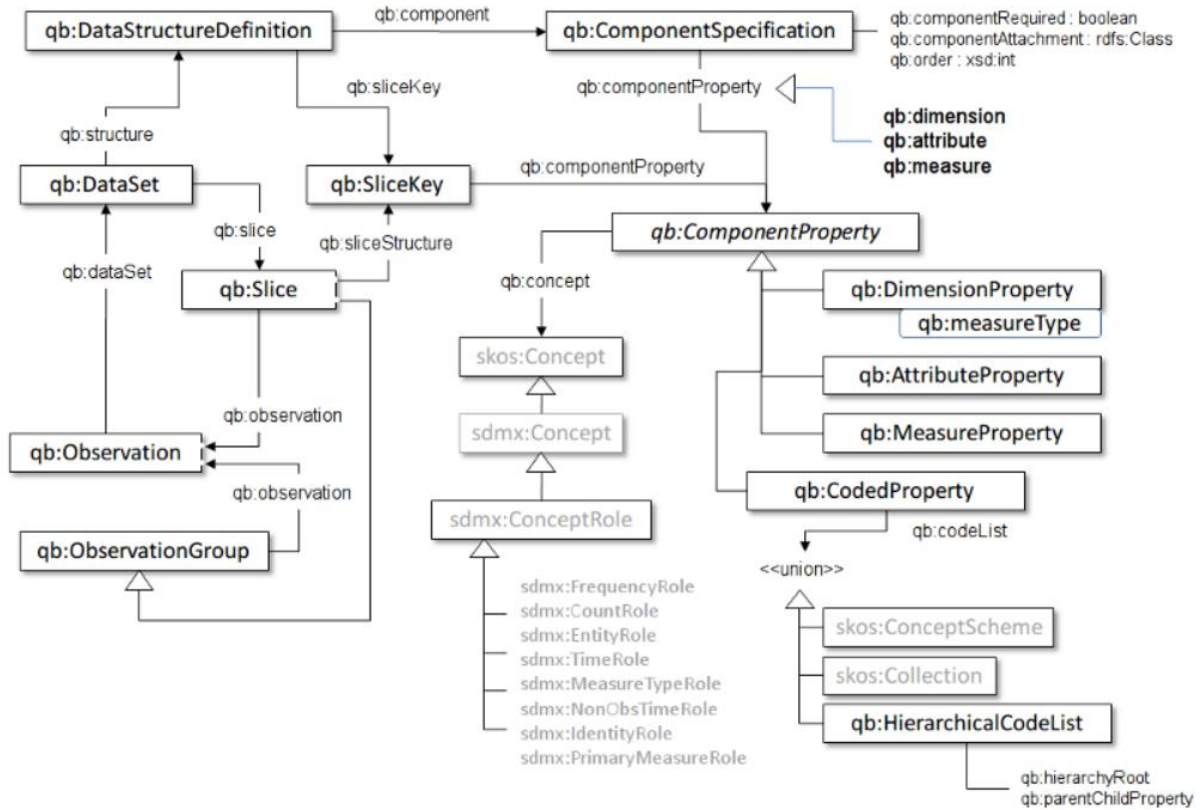
Πάνω στο πρότυπο SKOS έχει βασιστεί και ο EuroVoc, ο επίσημος πολύγλωσσος και επιστημονικός θησαυρός όρων της Ε.Ε., ο οποίος συμπεριλαμβάνει μεγάλο πλήθος εννοιών (concepts) σε μια οργανωμένη ιεραρχική και σημασιολογική δομή, καλύπτοντας τους τομείς δραστηριοτήτων της Ε.Ε. [11].

### 2.2.3 Λεξιλόγιο RDF Data Cube

Η αναπαράσταση μεμονωμένων εννοιών και ορολογιών καλύπτεται επαρκώς από το πρότυπο SKOS, ωστόσο η δημοσίευση πολύπλοκων, πολυδιάστατων στατιστικών δεδομένων - όπως οι πίνακες που διαθέτει η Eurostat - απαιτεί ένα πιο εξειδικευμένο σημασιολογικό εργαλείο. Για την κάλυψη αυτής της ανάγκης, το W3C έχει θεσπίσει το λεξιλόγιο RDF Data Cube Vocabulary [12] το οποίο είναι συνυφασμένο με το πρότυπο SDMX αποτελώντας την σημασιολογική του μεταφορά στον κόσμο των Linked Data. Συγκεκριμένα, η μοντελοποίηση γίνεται ως ένας πολυδιάστατος κύβος με άξονες αυτού τις διαστάσεις (dimensions), τα χαρακτηριστικά (attributes) και τις μετρήσεις (measures).

Για την πληρέστερη κατανόηση της μοντελοποίησης, τα συγκεκριμένα δομικά συστατικά του λεξιλογίου RDF Data Cube που καθορίζουν τον ρόλο κάθε πληροφορίας μέσα στον κύβο, είναι τα εξής [12]:

- **qb:DataSet** : Αντιπροσωπεύει το ίδιο το σύνολο δεδομένων (dataset).
- **qb:DataStructureDefinition** : Αποτελεί τη δομή του συνόλου (DSD), λειτουργώντας ως το «σχήμα» (schema) το οποίο επεξηγεί τον κύβο, δηλαδή ποιες διαστάσεις και μετρήσεις τον απαρτίζουν.
- **qb:component** : Συνδέει το DSD με τις διαστάσεις (dimensions) και το μετρήσιμο μέγεθος (measure).
- **qb:Observation** : Αντιπροσωπεύει κάθε μεμονωμένη παρατήρηση, δηλαδή ένα συγκεκριμένο στατιστικό νούμερο το οποίο αντιστοιχεί σε έναν μοναδικό συνδυασμό διαστάσεων.
- **sdmx-measure:obsValue** : Εκφράζει την αριθμητική τιμή της παρατήρησης/μέτρησης.



Σχήμα 2.5: Βασική δομή λεξιλογίου RDF Data Cube [12]

Στο πλαίσιο της εργασίας, η φιλοσοφία του RDF Data Cube αξιοποιείται κατά την ανάλυση των δομών δεδομένων (DSD) της Eurostat ενώ το SKOS χρησιμοποιείται από την εφαρμογή για τη σημασιολογική αναζήτηση κειμένων. Η λογική του Data Cube επιτρέπει στο σύστημα να κατανοεί τη δομή των στατιστικών πινάκων. Με αυτόν τον τρόπο, η εφαρμογή μπορεί να ανακτά δυναμικά τις διαστάσεις και να τις μετατρέπει σε πιθανά φίλτρα για τους χρήστες, προσφέροντας παράλληλα τη δυνατότητα εξαγωγής των δεδομένων σε έγκυρη, σημασιολογική μορφή RDF, έτοιμη προς διασύνδεση στον Σημασιολογικό Ιστό.

## 2.2.4 Λεξιλόγια Περιγραφής DCAT και DCTerms

Παρότι η χρήση του RDF Data Cube επαρκεί για την κάλυψη της στατιστικής δομής, κρίνεται αναγκαία η χρήση περιγραφικών λεξιλογίων για τα ίδια τα δεδομένα/σύνολα ως οντότητες. Για αυτόν τον σκοπό, χρησιμοποιούνται ευρέως τα λεξιλόγια Dublin Core (DCTerms) και Data Catalog Vocabulary (DCAT). Το DCTerms χρησιμοποιείται για την περιγραφή βασικών στοιχείων ενός εγγράφου όπως ο τίτλος (`dcterms: title`) και η περιγραφή (`dcterms:description`) [27]. Ενώ το DCAT στοχεύει κυρίως στην διευκόλυνση της διαλειτουργικότητας μεταξύ καταλόγων δεδομένων, επιτρέποντας την προσθήκη σημασιολογικών ετικετών και θεματικών κατηγοριών, όπως το `dc:keyword` [28].

## 2.3 Ανάκτηση Πληροφορίας

Ο τομέας της Ανάκτησης Πληροφορίας (Information Retrieval) πραγματεύεται τη δομή, την ανάλυση, την οργάνωση, την αποθήκευση και την αναζήτηση πληροφοριών, με κύριο στόχο τον εντοπισμό των πιο σχετικών εγγράφων ή δεδομένων που ανταποκρίνονται σε μια συγκεκριμένη ανάγκη πληροφόρησης

ενός χρήστη [13]. Σε αντίθεση με τις βάσεις δεδομένων που βασίζονται σε δομημένα ερωτήματα, τα συστήματα ανάκτησης πληροφορίας καλούνται να διαχειριστούν ημι-δομημένα ή αδόμητα κείμενα. Για να επιτευχθεί αυτό αποδοτικά, η διαδικασία διαχωρίζεται σε δύο κύριες φάσεις: τη δημιουργία ευρετηρίων (indexing) και τη βαθμολόγηση και κατάταξη των αποτελεσμάτων (scoring and ranking).

### 2.3.1 Δημιουργία Ευρετηρίων σε Σχεσιακή Βάση

Η απευθείας αναζήτηση λέξεων-κλειδιών στο σύνολο του κειμένου χιλιάδων εγγράφων έχει ως αποτέλεσμα την σημαντική αύξηση του χρόνου εκτέλεσης. Για τον λόγο αυτό, τα συστήματα IR βασίζονται στη δημιουργία ενός Αντεστραμμένου Ευρετηρίου (Inverted Index) το οποίο αντί να βασίζεται στη σχέση «έγγραφο προς λέξεις», χρησιμοποιεί τη σχέση «λέξη προς λίστα εγγράφων» [13].

Για την κατασκευή του ευρετηρίου, η ακατέργαστη πληροφορία πρέπει να υποστεί μια σειρά διαδοχικών σταδίων επεξεργασίας κειμένου [13]:

- 1) Λεκτική Ανάλυση (Tokenization): Το κείμενο σπάει σε μεμονωμένες λέξεις ή μικρότερες μονάδες (tokens) με αφαίρεση σημείων στίξης.
- 2) Απαλοιφή Λέξεων Διακοπής (Stop-words Removal): Αφαιρούνται λέξεις που εμφανίζονται πολύ συχνά αλλά δεν προσδίδουν κάποια σημασιολογική αξία όπως άρθρα και σύνδεσμοι.
- 3) Λημματοποίηση ή Αναγωγή στη Ρίζα (Stemming / Lemmatization): Αφαιρούνται οι πτώσεις στις λέξεις ώστε να μείνει μόνο το θέμα της καθεμιάς αποφεύγοντας αποκλίσεις στην αντιστοίχιση λόγω κλίσεων (π.χ. οι όροι «αγρότης», «αγροτική» ανάγονται στη ρίζα «αγροτ-»).

Στο πλαίσιο της εργασίας, η διαδικασία αυτή υλοποιείται μέσω του αρχείου index-builder.js που συλλέγει εκτός σύνδεσης (offline) τον κατάλογο ToC, τα μεταδεδομένα ESMS και τις σημασιολογικές ετικέτες του EuroVoc, εφαρμόζει τους παραπάνω κανόνες προεπεξεργασίας που αφορούν τη βασική λεκτική ανάλυση (Tokenization) και τον εντοπισμό στοιχειωδών λέξεων διακοπής (Stop-words Removal) και κατασκευάζει ένα τοπικό αντεστραμμένο ευρετήριο σε μορφή δομημένου αρχείου JSON και έναν φάκελο με όλα τα μεταδεδομένα ESMS σε JSON μορφή. Ωστόσο, τονίζεται ότι η Λημματοποίηση (Lemmatization/Stemming) δεν εφαρμόστηκε στην τωρινή εκδοχή της εφαρμογής και αποτελεί στόχο για μελλοντική επέκταση. Έπειτα, μετατρέπονται σε Σχεσιακή Βάση Δεδομένων (MySQL) ενώ η επικοινωνία με τον διακομιστή γίνεται μέσω Unix Socket. Με αυτόν τον τρόπο, όταν ο χρήστης αλληλεπιδρά με τη διεπαφή της εφαρμογής, η αναζήτηση εκτελείται πάνω στην δομημένη βάση δεδομένων, αποφεύγοντας την χρονοβόρα ζωντανή σάρωση της πλατφόρμας της Eurostat.

### 2.3.2 Μηχανισμοί Βαθμολόγησης και Κατάταξης

Μετά από την δημιουργία ευρετηρίου και τον εντοπισμό των εγγράφων πρέπει οι χρήστες να τα λαμβάνουν με σειρά σχετικότητας που πρέπει να υπολογίζεται. Η μαθηματική αποτίμηση αυτής της σχετικότητας βασίζεται συνήθως στο Μοντέλο Διανυσματικού Χώρου (Vector Space Model), όπου τόσο τα έγγραφα και το ερώτημα του χρήστη αναπαρίστανται ως διανύσματα σε έναν πολυδιάστατο χώρο όρων [14].

Ένας από τους πιο διαδεδομένους αλγόριθμους κατάταξης είναι ο BM25 (Best Matching 25), που αποτελεί εξέλιξη του TF-IDF (Term Frequency - Inverse Document Frequency) [13], με την διαφορά ότι στον BM25 εισάγει έναν μηχανισμό κορεσμού της συχνότητας (term frequency saturation) και κανονικοποίησης βάσει του μήκους του εγγράφου. Αυτό πρακτικά σημαίνει ότι η συνεχής επανάληψη μίας λέξης δεν συνεπάγεται την αύξηση της βαθμολογίας που λαμβάνει, ενώ τα μικρότερα έγγραφα ευνοούνται σε σχέση με τα μεγαλύτερα αν περιέχουν τον ίδιο αριθμό όρων αναζήτησης [13].

Στην αρχιτεκτονική της παρούσας εργασίας, υλοποιήθηκε ένας υβριδικός μηχανισμός βαθμολόγησης και κατάταξης που βασίζεται πάνω στον αλγόριθμο BM25, αξιοποιώντας τις εγγενείς δυνατότητες

αναζήτησης πλήρους κειμένου (Full-Text Search) της Σχεσιακής Βάσης Δεδομένων. Αρχικά, η αναζήτηση πραγματοποιείται στα εξής πεδία: στον τίτλο του dataset (title), στην θεματική ενότητα (theme), στο μοναδικό αναγνωριστικό (id), στις ετικέτες σημασιολογικού εμπλουτισμού (eurovoc) και στα εκτενή κείμενα μεταδεδομένων ESMS (description και concepts). Αντί να εφαρμόζεται ένα απλό άθροισμα για την εξαγωγή της βαθμολογίας, εφαρμόζεται μία στρατηγική ενίσχυσης πεδίων (field boosting). Τα διάφορα πολλαπλασιαστικά βάρη που αποδόθηκαν σε κάθε πεδίο, δόθηκαν έπειτα από διαδοχικές πειραματικές δοκιμές (trial and error), ώστε να επιτευχθεί η καλύτερη δυνατή συνάφεια αποτελεσμάτων. Συγκεκριμένα, αν ο όρος αναζήτησης εντοπιστεί στον κύριο τίτλο του dataset, το αποτέλεσμα λαμβάνει την υψηλότερη βαθμολογία με βάρος 100 ενώ αν εντοπιστεί στο θέμα ή στο αναγνωριστικό ID λαμβάνει βάρος 50 και 40 αντιστοίχως. Τέλος, αν εντοπιστεί στις σημασιολογικές ετικέτες του EuroVoc ή στα εκτενή κείμενα της περιγραφής ESMS, λαμβάνει βάρος 15 και 10 αντίστοιχα. Τη χαμηλότερη βαθμολογία με βάρος 10 λαμβάνει εξίσου και ο εντοπισμός στα Codelists. Το σύστημα αθροίζει τους επιμέρους δείκτες, κατατάσσει τα σύνολα δεδομένων και παρουσιάζει στον χρήστη μια ιεραρχημένη λίστα με τα πρώτα 50 κορυφαία datasets σε φθίνουσα σειρά βάσει score, διασφαλίζοντας ότι τα πιο συναφή έγγραφα θα εμφανιστούν στην κορυφή.

## 2.4 Σχετικές Εργασίες και Σύγχρονες Προσεγγίσεις

Τα Ανοιχτά Δεδομένα (Open Data) μαζί με τη ενσωμάτωση της σημασιολογικής τους αξίας αποτελούν βασικό ερευνητικό στόχο στα σύγχρονα πεδία της Πληροφορικής. Οι διεθνείς οργανισμοί προσπαθούν να μεταβούν από την δημοσιοποίηση απλών δεδομένων στην κοινοποίηση διασυνδεδεμένων, ενισχύοντας πρωτίστως την διαφάνεια και δευτερευόντως την διαλειτουργικότητα. Σε αυτήν την ενότητα, θα παρουσιασθούν πρόσφατες ερευνητικές προσεγγίσεις - κοινώς γνωστό ως State-of-the-Art - που συσχετίζονται με τους σκοπούς που θέλει να καλύψει η παρούσα εφαρμογή.

### 2.4.1 Προκλήσεις των Πυλών Ανοιχτών Κυβερνητικών Δεδομένων

Η πρακτική χρησιμότητα των δημοσιευμένων δεδομένων από κυβερνητικούς και ευρωπαϊκούς φορείς έχει αναλυθεί εκτενώς τα τελευταία χρόνια. Βάσει πρόσφατης βιβλιογραφικής ανασκόπησης [19], εξάγεται το συμπέρασμα της μειωμένης διαλειτουργικότητας που χαρακτηρίζει τα περισσότερα προγράμματα Open Government Data (OGD), δυσκολεύοντας ταυτοχρόνως και την κατανόησή τους από τους τελικούς χρήστες. Αντίστοιχα, η έλλειψη δυνατότητας επεξεργασίας και φιλτραρίσματος δημοσίων δεδομένων μέσω ισχυρών εργαλείων, οδηγεί σε αδύναμη εξαγωγή ουσιαστικής γνώσης [20].

Από τις παραπάνω έρευνες συμπεραίνεται ότι οι δημόσιοι φορείς – όπως είναι και αυτός της Eurostat – χρειάζονται ενδιάμεσα επίπεδα όπου θα απλοποιείται η πληροφορία. Εκεί ακριβώς στοχεύει και η παρούσα εφαρμογή, να λειτουργήσει ουσιαστικά ως «ενδιάμεσος», υπεύθυνος για το φιλτράρισμα και την αναδιοργάνωση της δομής του ToC.

### 2.4.2 Σημασιολογική Αναζήτηση σε Ανοιχτά Δεδομένα

Κατά την υλοποίηση ανάκτησης πληροφορίας, η χρήση της κλασικής αναζήτησης βάσει λέξεων-κλειδιών (keyword matching) κρίνεται ανεπαρκής και απαρχαιωμένος. Σε πρόσφατη έρευνα για το ερευνητικό έργο ODISSEI, αναλύθηκε η σημασιολογική αναζήτηση σε αποθετήρια ανοικτών δεδομένων και εξήχθη το παρακάτω συμπέρασμα. Ο αυτοματοποιημένος σημασιολογικός εμπλουτισμός των μεταδεδομένων μέσω εξόρυξης κειμένου και η διασύνδεσή τους με ελεγχόμενα λεξιλόγια συνθέτει την επίλυση του προβλήματος της ενοποιημένης αναζήτησης [21].

Στο Eurostat's Finder, ακολουθείται αυτή η διαδικασία με την αξιοποίηση κειμένου των εγχειριδίων μεταδεδομένων ESMS και την αυτοματοποιημένη διαδικασία αντιστοίχισης λέξεων του ευρωπαϊκού θησαυρού EuroVoc βάσει των παραπάνω κειμένων, επεκτείνοντας τις δυνατότητες της αναζήτησης.

### 2.4.3 Στατιστικά Δεδομένα και Λεξιλόγιο RDF Data Cube

Οι χειροκίνητες συνδέσεις είναι χαρακτηριστικό μειονέκτημα στις προσπάθειες κατασκευής γραφημάτων γνώσης για OGD, αφού απομονώνονται τα δημόσια datasets [22]. Η χρήση του λεξιλογίου RDF Data Cube προτείνεται ως ιδανική λύση για πολυδιάστατα στατιστικά δεδομένα όπως είναι αυτά της Eurostat.

Όμως η διαχείριση μεγάλων συνόλων RDF επιφέρει νέα προβλήματα. Αναλυτικότερα, για την μόνιμη αποθήκευση και εκτέλεση αναλυτικών ερωτημάτων πάνω σε τεράστια RDF δεδομένα, απαιτούνται αντίστοιχα τεράστιοι αποθηκευτικοί και υπολογιστικοί πόροι [23]. Ωστόσο, πρόσφατες πλατφόρμες όπως η FAIRLinked, επικεντρώνονται στην μετάβαση από την στατική διαχείριση και αποθήκευση των δεδομένων στην αντίστοιχη δυναμική, αξιοποιώντας τα θετικά του RDF Data Cube χωρίς τους αποθηκευτικούς και υπολογιστικούς περιορισμούς [24]. Αυτή η αρχιτεκτονική ακολουθήθηκε και στην παρούσα εφαρμογή.

### 2.4.4 Γενική Αποτίμηση της Εφαρμογής

Αξιολογώντας την σύγχρονη βιβλιογραφία, αναδεικνύεται η μονομερής προσέγγιση των περισσότερων είτε αποκλειστικά στην βελτίωση ποιότητας της αναζήτησης του χρήστη, είτε στη δημιουργία τριάδων RDF με τις μεγάλες απαιτήσεις τους.

Στο Eurostat's Finder γίνεται μία προσπάθεια σύμπτυξης αυτών των δύο προσεγγίσεων, δημιουργώντας μία υβριδική λύση που εκμεταλλεύεται τα πλεονεκτήματα της καθεμίας και αποφεύγει τους περιορισμούς της. Συνοπτικά, αποφεύγει την αποθήκευση ολόκληρης της βάσης της Eurostat σε RDF και προτιμά μια συμπιεσμένη Σχεσιακή Βάση σημασιολογικά εμπλουτισμένη από τα μεταδεδομένα. Ενώ η παραγωγή των στατιστικών σε μορφή RDF υλοποιείται δυναμικά, εκμηδενίζοντας τις απαιτήσεις υποδομών και προσφέροντας μια ελαφριά και αποδοτική προσέγγιση που συμμορφώνεται με τις αρχές των Ανοιχτών Δεδομένων [24].

Για τη σαφέστερη αποτύπωση των πλεονεκτημάτων της προτεινόμενης αρχιτεκτονικής έναντι των υφιστάμενων λύσεων που καταγράφονται στη σύγχρονη βιβλιογραφία, δημιουργήθηκε ο Πίνακας 2.1. Στον πίνακα αυτόν συνοψίζονται τα βασικά τεχνολογικά χαρακτηριστικά της εφαρμογής Eurostat's Finder, συγκρινόμενα με τις συμβατικές πύλες (όπως το Data Browser της Eurostat) και τα σύγχρονα ερευνητικά έργα (όπως ODISSEI και FAIRLinked).

Πίνακας 2.1: Συγκριτική Αξιολόγηση Χαρακτηριστικών Συστήματος και Υφιστάμενων Προσεγγίσεων

| Χαρακτηριστικό                | Eurostat DB | ODISSEI   | FAIRLinked | Eurostat's Finder |
|-------------------------------|-------------|-----------|------------|-------------------|
| Χρήση Μεταδεδομένων           | OXI         | NAI       | OXI        | NAI               |
| Εμπλουτισμός Λεξιλογίου       | OXI         | NAI       | OXI        | NAI               |
| Εξαγωγή σε RDF/Linked Data    | OXI         | ΜΕΡΙΚΩΣ   | NAI        | NAI               |
| Δυναμική Παραγωγή Δεδομένων   | N/A         | OXI       | NAI        | NAI               |
| Ενιαίο Δυναμικό Περιβάλλον UI | OXI         | ΕΞΑΡΤΑΤΑΙ | ΕΞΑΡΤΑΤΑΙ  | NAI               |

Από τον παραπάνω πίνακα γίνεται σαφές ότι η προσέγγιση του Eurostat's Finder δεν εστιάζει μονοδιάστατα σε ένα μόνο τεχνολογικό πεδίο, αλλά αποτελεί μια ολιστική, υβριδική λύση. Συγκεντρώνει τη σημασιολογική ευφυΐα έργων όπως το ODISSEI και την ελαφριά αρχιτεκτονική (on-the-fly) που έχουν πλατφόρμες όπως το FAIRLinked, ξεπερνώντας τους περιορισμούς των παραδοσιακών, συμβατικών πυλών ανοικτών δεδομένων.

## Κεφάλαιο 3ο: Αρχιτεκτονική Συστήματος

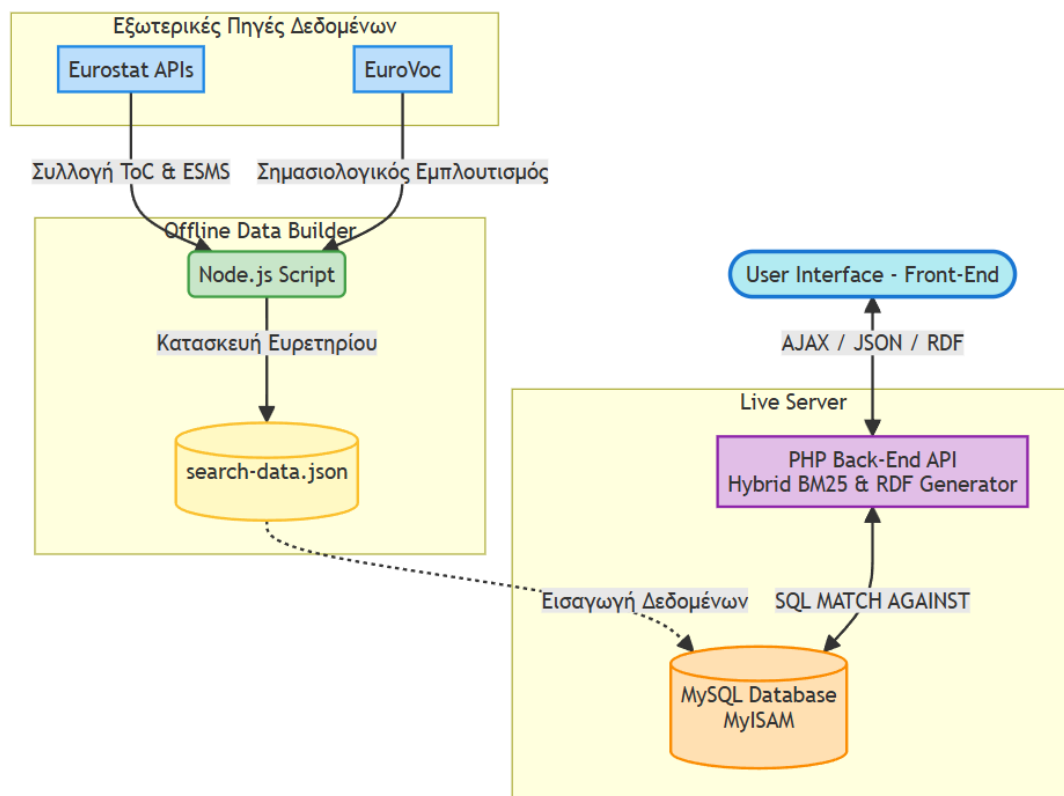
### 3.1 Γενική Επισκόπηση

Η αρχιτεκτονική του Eurostat's Finder χωρίζεται σε τρία βασικά επίπεδα: το τοπικό επίπεδο προετοιμασίας δεδομένων (Local Offline Data Builder), το επίπεδο εξυπηρέτησης (Live Back-end API) και το επίπεδο διεπαφής χρήστη (Front-end User Interface).

Αρχικά, ο τοπικός μηχανισμός προετοιμασίας αναλαμβάνει την συλλογή δεδομένων και μεταδεδομένων από την Eurostat και το EuroVoc ώστε να χτίσει δυναμικά το ενδιαμέσο αρχείο `search_data.json`, ένα ευρετήριο σημασιολογικά εμπλουτισμένο και με σωστή δομή, το οποίο μετατρέπεται σε μία Σχεσιακή Βάση Δεδομένων (MySQL).

Το live back-end επικοινωνεί απευθείας με την βάση δεδομένων μέσω Unix Sockets και αναλαμβάνει την εκτέλεση του αλγορίθμου κατάταξης και επιστροφής των συναφών αποτελεσμάτων ενώ χειρίζεται και την μετατροπή των δεδομένων στην σημασιολογική μορφή RDF.

Τέλος, το front-end απεικονίζει το User Interface (UI) της εφαρμογής και βάσει του προαναφερόμενου ευρετηρίου δίνει την δυνατότητα επιλογής συγκεκριμένου dataset και επιβολής δυναμικών στατιστικών φίλτρων στον χρήστη.



Σχήμα 3.1: Διάγραμμα Αρχιτεκτονικής Συστήματος

Για την καλύτερη κατανόηση της αλληλουχίας των διαδικασιών, η συνολική ροή της μεθοδολογίας μπορεί να συνοψιστεί στα ακόλουθα 11 διαδοχικά βήματα:

1. Ανάκτηση ToC από Eurostat
2. Εξαγωγή dataset ID, title, theme, parentTitle, ESMS URL
3. Ανάκτηση ESMS 3.1 (description) και 3.4 (concepts)
4. Ανάκτηση DSD και Codelists
5. Καθαρισμός/μείωση θορύβου
6. EuroVoc matching
7. Μετατροπή JSON σε MySQL
8. Αναζήτηση με weighted scoring
9. Επιλογή dataset και δυναμικά φίλτρα
10. Ανάκτηση δεδομένων από Eurostat API
11. Παραγωγή RDF DSD και RDF Observations

### 3.2 Μηχανισμός Συλλογής και Προεπεξεργασίας

Το αρχικό στάδιο της αρχιτεκτονικής ροής είναι η αυτοματοποιημένη συλλογή των στατιστικών δεδομένων μέσω των APIs της Eurostat και η μετέπειτα οργάνωσή αυτών των ετερογενών πληροφοριών. Λόγω του τεράστιου όγκου δεδομένων της Eurostat, ο μηχανισμός αυτός μπορεί να χωριστεί σε δύο μέρη: την ανάκτηση του ToC και των μεταδεδομένων και την εξόρυξη της δομής των δεδομένων

#### 3.2.1 Ανάκτηση Καταλόγου και Μεταδεδομένων

Η ανάκτηση του καταλόγου ToC σε μορφή xml γίνεται μέσω του API Catalogue της Eurostat και εξάγονται: τίτλος dataset, ID, θεματική ενότητα, γονική ενότητα και ο ESMS σύνδεσμος. Από τον τελευταίο εξάγονται οι ενότητες 3.1 και 3.4 που αφορούν την περιγραφή και τα concepts του εκάστοτε dataset αντίστοιχα. Για την σωστή διαχείριση του τεράστιου αυτού όγκου πληροφοριών, τα ESMS ID, description και concepts αποθηκεύονται ξεχωριστά από το ευρετήριο και μετέπειτα μετατρέπονται και αυτά σε πίνακα mySQL. Επειδή πολλά dataset παρόμοια θεματικά μεταξύ τους χρησιμοποιούν το ίδιο ESMS ID, συνεπώς καθώς μοιράζονται κοινή περιγραφή και έννοιες, η ξεχωριστή αποθήκευσή τους εξοικονομεί αποθηκευτικό χώρο. Η σύνδεση των δύο πινάκων επιτυγχάνεται με την ύπαρξη του FOREIGN KEY ESMS ID.

#### 3.2.2 Εξόρυξη Δομής Δεδομένων

Η ανάκτηση του ID δεν αρκεί για την επιμέρους ανάκτηση όλων των δεδομένων του εκάστοτε dataset. Το API SDMX 3.0 της Eurostat παρέχει τις κατάλληλες πληροφορίες μέσω εξειδικευμένων URLs με την κατάλληλη σειρά ώστε να ανακτηθούν τα Codelists, που ορίζουν τα σύνολα των αποδεκτών τιμών που μπορεί να λάβει η κάθε διάσταση ενός dataset, όπως και την αντιστοίχισή τους από κωδικές ονομασίες σε ανθρώπινα αναγνώσιμες τιμές (π.χ. «EL» αντιστοιχίζεται με «Greece»).

Η μεθοδολογική ροή της εξόρυξης της δομής δεδομένων ορίζεται από την εξής ακολουθία:

- 1) **Dataflow:** Με την χρήση του ID, που είναι μοναδικό για κάθε dataset, γίνεται εύρεση του εξίσου μοναδικού αναγνωριστικού δομής DSD ID.
- 2) **DSD Ανάλυση & Εύρεση Codelists:** Με την χρήση του DSD ID, εμφανίζονται οι διαστάσεις του συγκεκριμένου dataset και η κάθε μία παραπέμπει σε ένα συγκεκριμένο codelist συγκεκριμένης version (π.χ. 3.7).
- 3) **Αποκωδικοποίηση των Codelists:** Βάσει της έκδοσης του codelist, ανακτώνται τα πλήρη λεξιλόγια των διαστάσεων πετυχαίνοντας την αντιστοίχιση κωδικοποιημένων ονομασιών σε ανθρωπίνως αναγνώσιμες ετικέτες.

### 3.3 Έξυπνο Φιλτράρισμα και Μείωση Θορύβου

Μαζί με την ανάκτηση όλων των παραπάνω πληροφοριών, συγκεντρώνεται και πληροφορία που, αν και δομικά ορθή, εισάγει στατιστικό «θόρυβο» στις αναζητήσεις των χρηστών και καταλαμβάνει άσκοπα αποθηκευτικό χώρο. Η ανάγκη μείωσής του κρίνεται επιτακτική ώστε να μειωθεί ο όγκος του αρχείου και να βελτιστοποιηθεί η ποιότητα των αποτελεσμάτων.

Το σημαντικότερο βήμα μείωσης θορύβου αφορά την διαχείριση των διαστάσεων και των Codelists. Αρχικά, τα Codelists λειτουργούν ως πρότυπα λεξιλόγια, παρέχοντας τον κατάλογο των έγκυρων τιμών από τις οποίες αντλούν δεδομένα οι διαστάσεις. Όμως, οι πραγματικά ενεργές τιμές ανά κάθε dataset είναι συνήθως ένα υποσύνολο όλων των τιμών που ανακτήθηκαν από το codelist, με βασικό παράδειγμα την διάσταση «geo» («Geopolitical Entity») όπου γενικά περιλαμβάνει εκατοντάδες τιμές, όπως χώρες, πόλεις, και περιοχές χωρών, αλλά τα datasets που χρησιμοποιούν αυτήν την διάσταση αναφέρονται σε μία κατηγορία εξ αυτών και συχνά ακόμα και σε υποσύνολα αυτών, με τις υπόλοιπες τιμές να αποτελούν «θόρυβο». Η συγκεκριμένη διάσταση, όπως προαναφέρθηκε, δέχεται εκατοντάδες τιμές, αυξάνοντας δραματικά το μέγεθος του αρχείου. Επίσης, η εισαγωγή των ενεργών τιμών μπορεί να αύξανε τα «false positives» αφού σε άσχετα datasets όπου τυχαίνει να περιέχουν σε Codelists τους την Ελλάδα, θα θεωρούνταν σωστή αντιστοίχιση σε αναζήτηση του χρήστη για την λέξη «Ελλάδα»

Συνεπώς, κρίθηκε αναγκαίο οι τιμές των Codelists να μην εισάγονται στο ευρετήριο παρά μόνο να αναφέρονται οι κωδικές ονομασίες των διαστάσεων με τα αντίστοιχα ονόματα των Codelists τους π.χ. freq : Frequency.

### 3.4 Σημασιολογικός Εμπλουτισμός

Ο σημασιολογικός εμπλουτισμός αποτελεί κεντρικό πυλώνα της αρχιτεκτονικής του συστήματος. Τα datasets είναι χιλιάδες, μερικά αρκετά παρόμοια θεματικά, η ενδεχόμενη αναζήτηση λέξεων-κλειδίων μόνο από τους τίτλους τους δεν θα αρκούσε για να αποδώσει σωστά τις ερευνητικές ανάγκες των χρηστών. Εκτός από την ύπαρξη των ESMS που παρέχουν την περιγραφή και τις έννοιες του κάθε dataset, για να γεφυρωθεί το σημασιολογικό χάσμα, γίνεται χρήση του επίσημου θησαυρού λέξεων της EE EuroVoc. Αρχικά, το λεξικό έχει γίνει λήψη τοπικά και έπειτα από την σάρωση όλων των διαθέσιμων δεδομένων κειμένου κάθε dataset (Τίτλος, Θέμα, Μεταδεδομένα ESMS, Codelists), σε JSON ακόμα μορφή, αντιπαραβάλλεται κάθε όρος με τους επίσημους όρους του επίσημου θησαυρού της EuroVoc. Οι ταυτίσεις αποθηκεύονται και αυτές στη Σχεσιακή Βάση Δεδομένων (MySQL) σε ξεχωριστό πεδίο. Ενώ κατά την εξαγωγή των δεδομένων, οι όροι ενσωματώνονται στο γράφημα RDF χρησιμοποιώντας την ιδιότητα `dc:keyword`, προσθέτοντας θεματικό «βάρος» και επιτρέποντας στην μηχανή αναζήτησης να εντοπίζει datasets και μέσω σημασιολογικών συσχετισμών.

### 3.5 Σχεδιασμός Ενδιάμεσου Τοπικού Ευρετηρίου και Εισαγωγή στη Βάση Δεδομένων

Η αρχιτεκτονική του συστήματος υιοθετεί μια προσέγγιση τοπικής και εκτός σύνδεσης συλλογής πληροφοριών και μετέπειτα δομημένης αποθήκευσής τους, για να ξεπεραστούν οι περιορισμοί της ζωντανής επικοινωνίας με τα APIs της Eurostat σε κάθε ερώτημα χρήστη. Σε πρώτο στάδιο, ένα αυτοματοποιημένο script γραμμένο σε Node.js εκτελείται τοπικά και αναλαμβάνει την συλλογή στοιχείων όπως ο κατάλογος ToC, τα DSDs και τα ESMS μεταδεδομένα με εξόρυξη μόνο των δύο χρήσιμων ενότητων. Παράλληλα, εκτελεί και τον σημασιολογικό εμπλουτισμό της EuroVoc που αναφέρθηκε στην παραπάνω ενότητα. Ύστερα από το πέρας της διαδικασίας, τα αποτελέσματά της βρίσκονται συγκεντρωμένα στο ενδιάμεσο αρχείο `search_data.json`, δομημένα και «καθαρά» από άποψη θορύβου. Σε δεύτερο στάδιο, η μορφή JSON δεν κρίνεται αποδοτική για πολύπλοκα συστήματα

οπότε γίνεται μετατροπή του αρχείου σε Σχεσιακή Βάση Δεδομένων (MySQL). Δημιουργούνται δύο πίνακες : datasets και esms\_metadata, όπου ο πρώτος περιέχει τίτλο, θέμα, γονικό κόμβο, Codelists, ESMS ID και EuroVoc ετικέτες ενώ ο δεύτερος περιλαμβάνει ESMS ID, description και concepts πλήρους κειμένου.

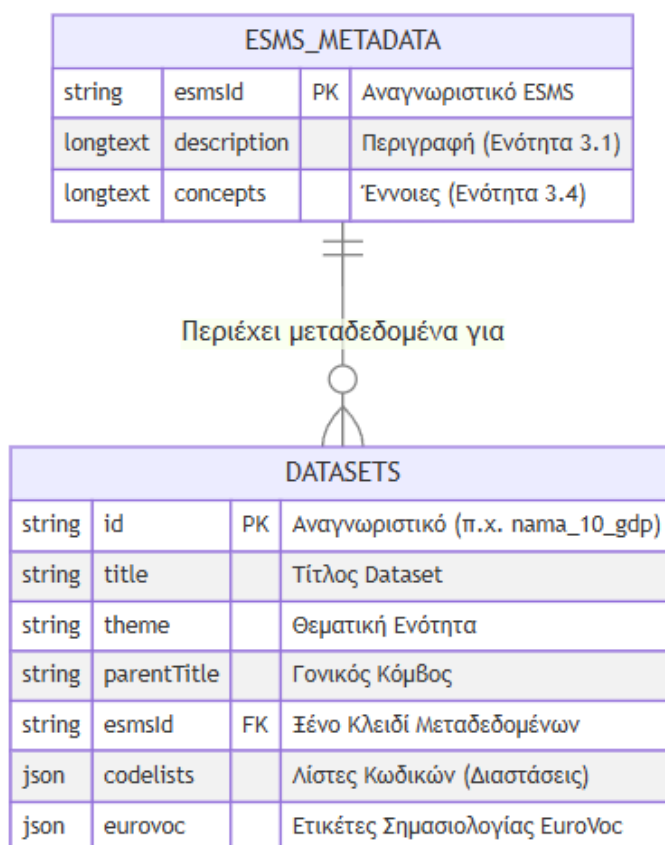
### 3.6 Σχεδιαστική Μοντελοποίηση Συστήματος

Για την πληρέστερη κατανόηση της αρχιτεκτονικής δομής της εφαρμογής, στην συγκεκριμένη ενότητα παρουσιάζονται διαγράμματα που αποτυπώνουν τις σχέσεις, που αναλύθηκαν παραπάνω, σε ένα συμπεριληπτικό πλαίσιο.

#### 3.6.1 Μοντέλο Οντοτήτων-Σχέσεων

Η σωστή αξιοποίηση του αποθηκευτικού χώρου και η αποφυγή πλεοναστικών δεδομένων αποτέλεσε κύριο σχεδιαστικό άξονα. Όπως παρουσιάζεται στο παρακάτω διάγραμμα, η Βάση Δεδομένων σχεδιάστηκε βάσει κανόνων κανονικοποίησης και χωρίστηκε σε δύο κύριους πίνακες: datasets και esms\_metadata.

## Διάγραμμα Οντοτήτων-Συσχετίσεων

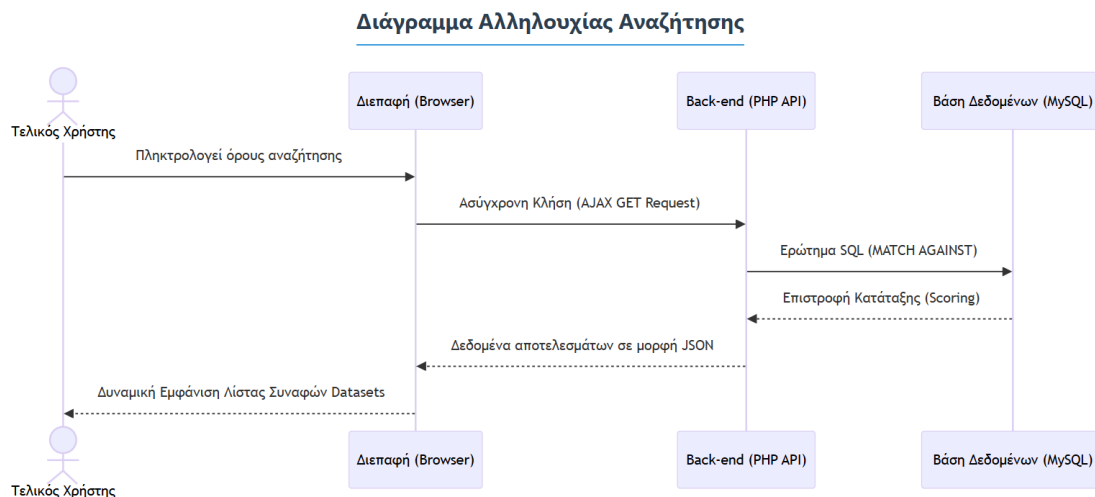


Σχήμα 3.2: Διάγραμμα Οντοτήτων-Συσχετίσεων

Η συσχέτιση των δύο πινάκων πραγματοποιείται μέσω του ξένου κλειδιού (Foreign Key- FK) esmsId. Με αυτόν τον τρόπο τα πιθανά διπλότυπα, λόγω της πολλαπλής χρήσης του ίδιου ESMS από διαφορετικά datasets, εξαφανίζονται.

### 3.6.2 Διάγραμμα Αλληλουχίας Αναζήτησης

Το διάγραμμα αλληλουχίας αναζήτησης απεικονίζει την συμπεριφορά της εφαρμογής, από το στάδιο της ηλεκτρολόγησης του χρήστη στην μπάρα αναζήτησης έως την εμφάνιση του dataset που έχει επιλέξει. Σκοπός του διαγράμματος είναι να τονίσει την δυναμική συμπεριφορά του συστήματος παρέχοντας φίλτρα και μετέπειτα αποτελέσματα σε πραγματικό χρόνο, χωρίς την ανάγκη επαναφόρτωσης σελίδας.



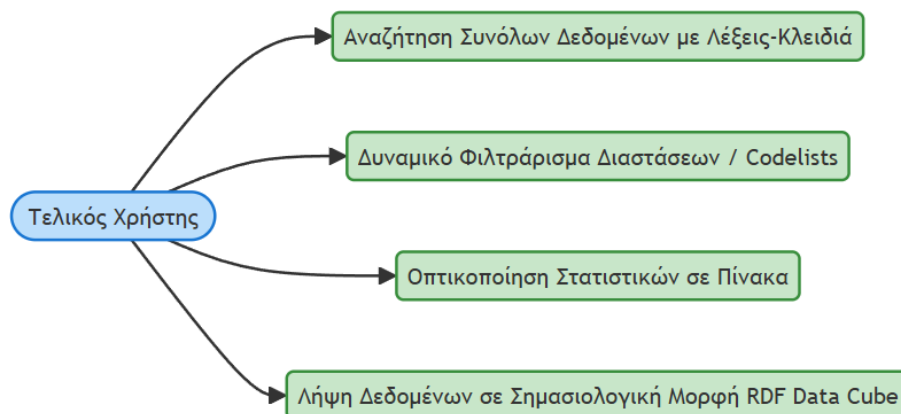
Σχήμα 3.3: Διάγραμμα Αλληλουχίας Αναζήτησης

Συνοπτικά, ο τελικός χρήστης πληκτρολογεί όρους αναζήτησης, η διεπαφή χρήστη (UI) αρχίζει μία ασύγχρονη κλήση AJAX προς το API και εκείνο με την σειρά του το μεταφράζει καταλλήλως και εκτελεί αντίστοιχο ερώτημα στην βάση δεδομένων. Η βάση κατατάσσει τα βάρη και ταξινομεί τα αποτελέσματα σε φθίνουσα σειρά. Τα αποτελέσματα μεταφέρονται αντίστροφα τώρα, από το back-end στο front-end σε μορφή JSON και εμφανίζονται τελικώς στην διεπαφή του χρήστη.

### 3.6.3 Διάγραμμα Ενεργειών και Δυνατοτήτων Χρήστη

Τέλος, το διάγραμμα ενεργειών και δυνατοτήτων χρήστη υλοποιήθηκε για να συνοψίσει τις δυνατότητες που προσφέρει το πρόγραμμα στον τελικό χρήστη.

### Διάγραμμα Ενεργειών Χρήστη



Σχήμα 3.4: Διάγραμμα Ενεργειών Χρήστη

Όπως αποτυπώνεται στο διάγραμμα, στον τελικό χρήστη παρέχεται η δυνατότητα επιλογής/αναζήτησης dataset βάσει εισαγωγής όρων, ενώ μετέπειτα μπορεί να φιλτράρει τα δεδομένα που θα ανακτήσει βάσει συγκεκριμένων φίλτρων/διαστάσεων που υποστηρίζει το κάθε dataset. Με το πέρας της διαδικασίας «φιλτραρίσματος», τα δεδομένα αναπαράστώνται σε αναγνώσιμη μορφή πίνακα με στατιστικές τιμές, ενώ δίνεται η δυνατότητα λήψης των αποτελεσμάτων σε μορφή RDF Data Cube, έτοιμων για χρήση στον Σημασιολογικό Ιστό.

## Κεφάλαιο 4ο: Υλοποίηση Συστήματος

### 4.1 Τεχνολογίες και Εργαλεία Υλοποίησης

Για την υλοποίηση του Eurostat's Finder απαιτήθηκαν διάφορες γλώσσες προγραμματισμού, διαδικτυακά πρότυπα και επιμέρους τεχνολογικοί μηχανισμοί προκειμένου να γεφυρωθούν τα χάσματα μεταξύ της αρχικής ακατέργαστης πληροφορίας μέχρι της τελικής που φτάνει στον χρήστη. Τα παραπάνω εργαλεία μπορούν να χωριστούν, βάσει χρήσης, σε τρεις κατηγορίες: τεχνολογίες υποβάθρου και διαχείρισης δεδομένων, τεχνολογίες διεπαφής χρήστη και εξωτερικά πρότυπα.

#### 4.1.1 Τεχνολογίες Υποβάθρου και Διαχείριση Δεδομένων

- **Node.js:** Αποκλειστικά τοπική και off-line χρήση για την αρχική ανάκτηση και μετέπειτα την προετοιμασία των δεδομένων. Λόγω της ασύγχρονης αρχιτεκτονικής του, αποτέλεσε τον κατάλληλο υποψήφιο για την ταχύτατη παράλληλη ανάκτηση εγγράφων μεγάλου όγκου της Eurostat. Κυρίως παρελάμβανε αρχεία XML και HTML και παρήγαγε αρχεία τύπου JSON. Χαρακτηριστικό παράδειγμα το τοπικό αρχείο index-builder.js, το οποίο έλαβε και επεξεργάστηκε δεδομένα, έκανε τις απαραίτητες ταυτίσεις με τον θησαυρό της EuroVoc και εν τέλει παρήγαγε το ενδιάμεσο αρχείο search-data.json.
- **PHP:** Κύρια γλώσσα εκτέλεσης στον διακομιστή παραγωγής. Αναλαμβάνει τον ρόλο της επικοινωνίας μεταξύ της Βάσης Δεδομένων, την προσομοίωση του αλγορίθμου βαθμολόγησης και την on-the-fly κατασκευή των αρχείων RDF. Επίσης, αναλαμβάνει τον ρόλο του ενδιάμεσου (proxy) για την αποφυγή σφαλμάτων διασταυρούμενης προέλευσης (CORS) κατά την κλήση των Eurostat APIs.
- **MySQL:** Επιλέχθηκε για γλώσσα της Βάσης Δεδομένων λόγω της εγγενούς υποστήριξής της σε αναζητήσεις πλήρους κειμένου. Ο πίνακας με τα ESMS metadata έχει δύο πεδία description και concepts όπου το καθένα ενίοτε περιλαμβάνει χιλιάδες χαρακτήρες. Επίσης, στρατηγικής σημασίας ήταν η χρήση της μηχανής αποθήκευσης MyISAM, που προσφέρει υψηλές ταχύτητες ανάγνωσης και ισχυρή συμπίεση, διευκολύνοντας την αποθήκευση τεράστιων πεδίων κειμένου όπως τα παραπάνω χωρίς να παραβιάζονται τα όρια (quotas) του διακομιστή.

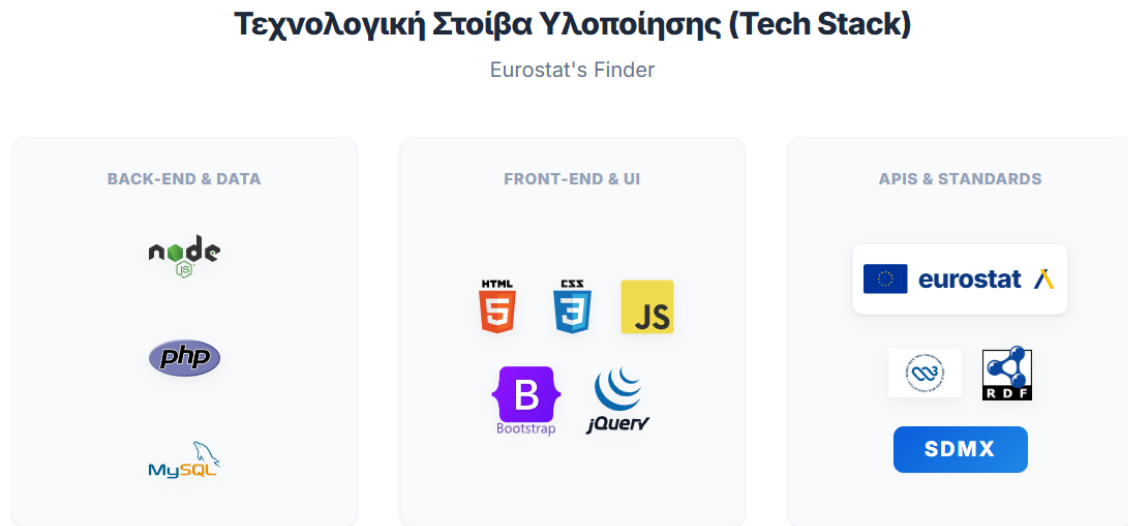
#### 4.1.2 Τεχνολογίες Διεπαφής Χρήστη

- **HTML5 & CSS3 (Bootstrap Framework):** Η δομή της σελίδας σχεδιάστηκε σε HTML5, ενώ σχεδιαστικά για μορφοποίηση και προσαρμοστικότητα σε διαφορετικά πλαίσια και συσκευές χρησιμοποιήθηκε το πλαίσιο Bootstrap, ένα σύγχρονο περιβάλλον με έτοιμα στοιχεία.
- **JavaScript:** Οποιαδήποτε δυναμική αλλαγή εμφανίζεται στην σελίδα του χρήστη υλοποιήθηκε με την JavaScript. Ιδιαίτερη αναφορά αξίζει να δοθεί στην βιβλιοθήκη jQuery που αξιοποιήθηκε για την απλοποίηση των ασύγχρονων κλήσεων \$.ajax και \$.get, επιτρέποντας την δυναμική ανανέωση της σελίδας χωρίς την απαίτηση επαναφόρτωσής της. Χαρακτηριστικά παραδείγματα η ανανέωση των φίλτρων και η ανάκτηση των αποτελεσμάτων.

#### 4.1.3 Εξωτερικά Πρότυπα

- **Eurostat APIs:** Η μόνη πηγή πρωτογενών δεδομένων και μεταδεδομένων για την Eurostat γίνεται μέσω των δικών της APIs. Κυρίως, χρησιμοποιήθηκε το SDMX 3.0 API που δίνει παραπάνω σχεσιακές και στατιστικές πληροφορίες από τα άλλα. Το Catalogue API χρησιμοποιήθηκε για την ανάκτηση του ToC σε XML μορφή ενώ το Statistics API βοήθησε μόνο στην εύρεση των πραγματικών, ενεργών τιμών των φίλτρων.
- **W3C Standards (RDF):** Η εξαγωγή των δεδομένων έγινε με το μοντέλο RDF και συγκεκριμένα σε σύνταξη Turtle (.ttl) λόγω της ανθρώπινης αναγνωσιμότητάς της και από την συμπίεση που προσφέρει.
- **RDF Data Cube Vocabulary & SKOS:** Τα επίσημα λεξιλόγια για την αναπαράσταση στατιστικών κύβων και θησαυρών λέξεων (EuroVoc), αντίστοιχα. Η χρήση τους καθιστά τα

παραγόμενα αποτελέσματα διαλειτουργικά στον ιστό των Συνδεδεμένων Ανοικτών Δεδομένων (Linked Open Data).



Σχήμα 4.1: Τεχνολογική Στοιβά Υλοποίησης της Eurostat's Finder

## 4.2 Κατασκευή και Δόμηση των Πινάκων

Το πρώτο βήμα για την κατασκευή πινάκων ήταν η ανάπτυξη του τοπικού script `index-builder.js` σε περιβάλλον Node.js, το οποίο ανακτά το αρχείο ToC μέσω του Catalogue API και εξάγει τα βασικά αναγνωριστικά των datasets όπως ο τίτλος, το αναγνωριστικό id και τον σύνδεσμο για τα μεταδεδομένα (ESMS URLs) ενώ αξιοποιώντας την δομή του ToC εξάγονται και ο γονικός κόμβος και το θέμα των datasets όπως φαίνεται και στο σχήμα 4.2.

```

67  const datasets = [];
68  const leafRegex = /<nt:leaf\b(?:[^\s]*)?>([\s\S]*?)<\nt:leaf>/g;
69  let match;
70  while ((match = leafRegex.exec(xml)) !== null) {
71      const attrs = match[1];
72      const content = match[2];
73      const typeMatch = attrs.match(/type="([^\"]+)"/);
74      if (!typeMatch || typeMatch[1] !== 'dataset') continue;
75      const codeMatch = content.match(/<nt:code>([^\s]+)<\nt:code>/);
76      if (!codeMatch) continue;
77      const code = codeMatch[1].trim();
78      const titleMatch = content.match(/<nt:title\s+language="en">([^\s]+)<\nt:title>/);
79      const title = titleMatch ? titleMatch[1].trim() : code;
80      const esmsMatch = content.match(/<nt:metadata\s+format="html">([^\s]+)<\nt:metadata>/);
81      const esmsUrl = esmsMatch ? esmsMatch[1].trim() : null;
82      let esmsId = null;
83      if (esmsUrl) {
84          const m = esmsUrl.match(/\/([^\s]+)_esms\.htm$/);
85          if (m) esmsId = m[1];
86      }
87      const pos = match.index;
88      const xmlBefore = xml.substring(0, pos);
89      const branchTitles = [...xmlBefore.matchAll(/<nt:branch>[\s\S]*?<nt:title language="en">([^\s]+)<\nt:title>/g)];
90      const parentTitle = branchTitles.length > 0 ? branchTitles[branchTitles.length - 1][1].trim() : '';
91      const allTitles = [...xmlBefore.matchAll(/<nt:title language="en">([^\s]+)<\nt:title>/g)];
92      const theme = allTitles.length > 1 ? allTitles[1][1].trim() : parentTitle;
93      datasets.push({ id: code, title, theme, parentTitle, esmsId, esmsUrl });
94  }
95  return datasets;

```

Σχήμα 4.2: Εξαγωγή βασικών αναγνωριστικών από ToC

Ιδιαίτερη προσοχή δόθηκε στην μετέπειτα αξιοποίηση των ESMS URLs με τον τεράστιο όγκο δεδομένων. Μέσω της βιβλιοθήκης node-html-parser, αναπτύχθηκε μηχανισμός σάρωσης του DOM (Document Object Model) που σαρώνει δυναμικά τις ESMS σελίδες και διαλέγει μόνο τις ενότητες 3.1, που περιέχει την περιγραφή του dataset και 3.4, που περιέχει τα concepts, απορρίπτοντας το υπόλοιπο κείμενο (Σχήμα 4.3).

```

132 function parseEsmsHtml(html, esmsId) {
133     const root = parseHtml(html);
134     const result = { description: '', concepts: '', esmsId };
135     root.querySelectorAll('h3, h2').forEach(h => {
136         const text = h.textContent.trim();
137         let field = null;
138         if (text.startsWith('3.1')) field = 'description';
139         else if (text.startsWith('3.4')) field = 'concepts';
140         if (!field) return;
141         let content = '';
142         let next = h.nextElementSibling;
143         while (next && !['H2', 'H3'].includes(next.tagName)) {
144             content += ' ' + next.textContent;
145             next = next.nextElementSibling;
146         }
147         result[field] = content.replace(/\s+/g, ' ').trim();
148     });
149     return result;
150 }

```

Σχήμα 4.3: Συνάρτηση ανάκτησης description και concepts

Για την ενίσχυση του σημασιολογικού εμπλουτισμού, εκτελείται σε Node.js το αρχείο fetch\_eurovoc.js πριν από το index-builder.js, αποσκοπώντας στην λήψη όλων των prefLabels και altLabels του θησαυρού της EuroVoc και την αποθήκευσή τους σε JSON μορφή στο eurovoc-labels.json. Έπειτα, με την εκτέλεση του μηχανισμού index-builder, συνενώνεται σε μία μεταβλητή όλη η διαθέσιμη πληροφορία ενός dataset (τίτλος, θέμα, διαστάσεις/Codelists, κείμενα ESMS) η οποία περνάει ως παράμετρος από την συνάρτηση matchEurovoc(). Με αυτόν τον τρόπο, σαρώνεται όλο το κείμενο και αντιστοιχίζονται τα labels του με τα αποθηκευμένα labels της EuroVoc. Προκειμένου να εξασφαλιστεί ο αποδοτικός σημασιολογικός εμπλουτισμός, η συνάρτηση ταξινομεί σε φθίνουσα σειρά βάσει μήκους χαρακτήρων ώστε να δοθεί προτεραιότητα σε μεγαλύτερους όρους που συνήθως είναι πιο εξειδικευμένοι. Τέλος, η συνάρτηση επιστρέφει τα 15 επικρατέστερα labels που αποθηκεύονται στο πεδίο eurovoc σε μορφή JSON.

```

239  async function buildIndex() {
265      }
266      const dims = await fetchDimensions(ds.id);
267      const fullText = [
268          ds.title, ds.theme, ds.parentTitle,
269          esms.description, esms.concepts,
270          Object.values(dims.codelists).map(cl => cl.name).join(' '),
271      ].join(' ');
272      const eurovoc = matchEurovoc(fullText);

```

Σχήμα 4.4: Συνένωση όλων των πεδίων στην μεταβλητή fullText

```

28  function loadEurovocLabels() {
29      if (fs.existsSync(EUROVOC_CACHE)) {
30          const labels = JSON.parse(fs.readFileSync(EUROVOC_CACHE, 'utf8'));
31          log(`EuroVoc: φορτώθηκαν ${labels.length} labels από cache.`);
32          return labels;
33      }
34      log(`EuroVoc: δεν βρέθηκε cache! Τρέξε πρώτα: node fetch_eurovoc.js`);
35      return [
36          'agriculture', 'fisheries', 'food safety', 'environment', 'climate change',
37          'energy', 'transport', 'economy', 'finance', 'gross domestic product',
38          'employment', 'unemployment', 'population', 'health', 'education'
39      ];
40  }
41
42  const EUROVOC_LABELS = loadEurovocLabels();
43
44  function matchEurovoc(text) {
45      const lower = text.toLowerCase();
46      const matched = EUROVOC_LABELS.filter(label => lower.includes(label));
47      matched.sort((a, b) => b.length - a.length);
48      return matched.slice(0, 15);
49  }

```

Σχήμα 4.5: Συνάρτηση ταύτισης και εξαγωγής όρων του EuroVoc

Με το πέρας της διαδικασίας index-builder.js, ολοκληρώνεται το αρχείο search-data.json, το ενδιαμέσο JSON αρχείο, που χρησιμοποιήθηκε για την κατασκευή της Σχεσιακής Βάσης Δεδομένων MySQL eurostat.db. Η μετατροπή από JSON σε MySQL έγινε από το PHP αρχείο mysql-converter.php το οποίο φτιάχνει τους δύο πίνακες datasets και esms\_metadata και ενημερώνει κατάλληλα τις τιμές τους από το αρχείο search-data.json. Ο σχεδιασμός του σχεσιακού σχήματος βασίστηκε στις αρχές της κανονικοποίησης, διαχωρίζοντας τα βασικά μεταδεδομένα σε δύο πίνακες, όπως φαίνεται στο Σχήμα 4.6, έχοντας κοινό πεδίο το esmsId. Η χρήση αυτού ως Ξένο Κλειδί ( FOREIGN KEY) εξαλείφει τον πλεονασμό των δεδομένων και συνεπώς μειώνει το μέγεθος του αρχείου αφού πολλά datasets εκφράζονται από το ίδιο ESMS ID. Στρατηγική επιλογή ωστόσο υπήρξε και η χρήση της μηχανής MyISAM που παρότι υστερεί σε σχέση με την InnoDB όσον αφορά σε υποστήριξη συναλλαγών και κλειδωμάτων σε επίπεδο γραμμής, υπερτερεί στην συμπίεση μεγάλων κειμένων λόγω της απουσίας περιττών undo logs και transaction overheads, επιτρέποντας την δημιουργία ευρετηρίων μεγάλου κειμένου όπως είναι τα δύο πεδία description και concepts του πίνακα esms\_metadata.

```

3  $socket = '/home/student/iee/2019/iee2019060/mysql/run/mysql.sock';
4  try {
5      $dsn = "mysql:unix_socket=$socket;charset=utf8mb4";// ΣΥΝΔΕΣΗ ΣΤΟΝ MySQL Server
6      $pdo = new PDO($dsn, $DB_USER, $DB_PASS);
7      $pdo->setAttribute(PDO::ATTR_ERRMODE, PDO::ERRMODE_EXCEPTION);
8      $pdo->exec("CREATE DATABASE IF NOT EXISTS `{$DB_NAME}` CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci");
9      $pdo->exec("USE `{$DB_NAME}`");
10     $pdo->exec("DROP TABLE IF EXISTS datasets;"); // ΚΑΘΑΡΙΣΜΟΣ ΠΑΛΙΩΝ ΠΙΝΑΚΩΝ ΣΕ ΠΕΡΙΠΤΩΣΗ ΠΡΟΗΓΟΥΜΕΝΟΥ CRASH
11     $pdo->exec("DROP TABLE IF EXISTS esms_metadata;");
12     $pdo->exec("CREATE TABLE esms_metadata (
13         esmsId VARCHAR(100) PRIMARY KEY,
14         description LONGTEXT,
15         concepts LONGTEXT
16     ) ENGINE=MyISAM DEFAULT CHARSET=utf8mb4;");
17     $pdo->exec("CREATE TABLE datasets (
18         id VARCHAR(100) PRIMARY KEY,
19         title TEXT,
20         theme TEXT,
21         parentTitle TEXT,
22         esmsId VARCHAR(100),
23         codelists JSON,
24         eurovoc JSON
25     ) ENGINE=MyISAM DEFAULT CHARSET=utf8mb4;");
26     $pdo->exec("ALTER TABLE datasets ADD FULLTEXT INDEX ft_title (title);");
27     $pdo->exec("ALTER TABLE datasets ADD FULLTEXT INDEX ft_theme (theme);");
28     $pdo->exec("ALTER TABLE esms_metadata ADD FULLTEXT INDEX ft_esms (description, concepts);");
29

```

Σχήμα 4.6: Δημιουργία πινάκων στην mysql-converter.php

### 4.3 Μηχανή Αναζήτησης

Η αποδοτική και γρήγορη συσχέτιση όρων αναζητήσεως με πιθανά datasets είναι σχέση-βαρόμετρο για την εφαρμογή. Η μηχανή αναζήτησης, που υλοποιείται μέσω του αρχείου `search_datasets.php`, αναλαμβάνει να συγκρίνει την συμβολοσειρά που εισάγει ο χρήστης με τα υπονήφια datasets και μετέπειτα να εμφανίσει τα πιο συναφή. Για την υλοποίηση αυτής, αξιοποιήθηκε η ισχύς των Ευρετηρίων μεγάλων κειμένων της MySQL, σε συνδυασμό με ερωτήματα μερικής ταύτισης (LIKE). Ο αλγόριθμος BM25 με την λογική της κατανομής βαρών στα πεδία χρησιμοποιήθηκε για την αρχική αντιστοίχιση πεδίων σε σκορ και την μετέπειτα πρόσθεση όλων των βαρών σε ένα συνολικό σκορ. Κατά την επεξεργασία των ερωτημάτων πολλών λέξεων, το σύστημα εκτελεί λεκτική ανάλυση και εφαρμόζει τη στρατηγική Cross-Field AND Matching. Συνεπώς, αυτό σημαίνει ότι πρέπει αναγκαστικά όλοι οι όροι της αναζήτησης να βρίσκονται στο ευρετήριο τουλάχιστον μία φορά, έστω και διασκορπισμένοι σε διαφορετικά πεδία, για να θεωρηθεί σχετικό το dataset και μετέπειτα αν συγκεντρώσει μεγάλο score να εμφανιστεί στην λίστα των επικρατέστερων. Αντιθέτως, αν έστω ένας όρος δεν συμπεριλαμβάνεται σε κανένα πεδίο, τότε το dataset λαμβάνει μηδενικό score, κρίνεται δηλαδή μη σχετικό με την συγκεκριμένη αναζήτηση. Η προσέγγιση αυτή σχεδιάστηκε αποσκοπώντας στην ελαχιστοποίηση των ψευδώς θετικών (false positives) αποτελεσμάτων, αυξάνοντας αισθητά την ακρίβεια (precision) της αναζήτησης. Ωστόσο, η απόλυτη αυτή λύση μπορεί να οδηγήσει σε ορισμένες περιπτώσεις σε μείωση της ανάκλησης (recall), αποκλείοντας datasets που αν και είναι σχετικά, δεν περιέχουν ολόκληρο το σύνολο των λέξεων του ερωτήματος αναζήτησης. Στις περιπτώσεις που υπάρχουν όλοι οι όροι της αναζήτησης στο ευρετήριο, ο αλγόριθμος πολλαπλασιάζει την εύρεση των όρων με τα αντίστοιχα βάρη των πεδίων στα οποία εντοπίστηκαν. Τα βάρη –όπως αναλύθηκαν στο θεωρητικό υπόβαθρο- είναι τα εξής :

- Ταίριασμα στον **Τίτλο** λαμβάνει πολλαπλασιαστή ίσο με 100, η πιο δυνατή συσχέτιση που μπορεί να επιτευχθεί
- Ταίριασμα στο **Θέμα** λαμβάνει πολλαπλασιαστή ίσο με 50, δυνατή συσχέτιση.

- Άμεση αναζήτηση του αναγνωριστικού **Κωδικού** id λαμβάνει βάρος ίσο με 40.
- Ταίριασμα στο σύνολο λέξεων της **EuroVoc** επιβραβεύεται με βάρος ίσο με 15, αξιοποιώντας τον σημασιολογικό εμπλουτισμό σε πιθανά συνώνυμα και γενικότερες έννοιες. Λόγω της γενικότητάς του όμως λαμβάνει σημαντική μείωση πολλαπλασιαστή από τα προηγούμενα για την αποφυγή πιθανών ψευδών θετικών (false positives).
- Τέλος, το ταίριασμα στα εκτενή κείμενα του **Πίνακα esms\_metadata** (description, concepts), καθώς και εντός των ονομάτων των **Codelists**, επιβραβεύονται με πολλαπλασιαστή ίσο με 10, τον μικρότερο όλων για τους ίδιους λόγους που προαναφέρθηκαν παραπάνω. Η διαφορά του βάρους σε σχέση με την EuroVoc δικαιολογείται από τα εκτενή κείμενα του πίνακα μεταδεδομένων που αυξάνει την πιθανότητα ψευδών θετικών ταιριασμάτων.

Τέλος, αξίζει να σημειωθεί ότι τα παραπάνω βάρη δεν αποτελούν απόλυτους αριθμούς, αλλά πολλαπλασιαστές. Η συνάρτηση MATCH() AGAINST() της MySQL επιστρέφει από μόνη της έναν δεκαδικό αριθμό σχετικότητας ο οποίος εξαρτάται από τον αριθμό εμφάνισης των όρων, το μήκος του κειμένου και τη σπανιότητα της λέξης. Αυτός ο αρχικός δεκαδικός αριθμός πολλαπλασιάζεται στη συνέχεια με το αντίστοιχο βάρος του πεδίου. Συνεπώς, εάν ένας όρος βρεθεί στον τίτλο με αρχικό σκορ 2.5 από τη MySQL, το τελικό σκορ μόνο από το πεδίο του τίτλου θα εκτοξευθεί στο 250 (2.5 \* 100). Αθροίζοντας τα επιμέρους γινόμενα όλων των πεδίων, προκύπτει το τελικό score του εκάστοτε dataset. Η μηχανή αναζήτησης επιστρέφει στον χρήστη τα πενήντα πρώτα σε συνολικό σκορ σε φθίνουσα σειρά σε μορφή JSON.

```

37  try {
38      $sql = "
39          SELECT
40              d.id AS code,
41              d.title,
42              (
43                  (MATCH(d.title) AGAINST(:q) * 100) +
44                  (MATCH(d.theme) AGAINST(:q) * 50) +
45                  (IFNULL(MATCH(e.description, e.concepts) AGAINST(:q), 0) * 10) +
46                  (CASE WHEN d.eurovoc LIKE :likeQ THEN 15 ELSE 0 END) +
47                  (CASE WHEN d.codelists LIKE :likeQ THEN 10 ELSE 0 END) +
48                  (CASE WHEN d.id LIKE :likeQ THEN 40 ELSE 0 END)
49              ) AS score
50          FROM datasets d
51          LEFT JOIN esms_metadata e ON d.esmsId = e.esmsId
52          WHERE
53              MATCH(d.title) AGAINST(:q)
54              OR MATCH(d.theme) AGAINST(:q)
55              OR MATCH(e.description, e.concepts) AGAINST(:q)
56              OR d.eurovoc LIKE :likeQ
57              OR d.codelists LIKE :likeQ
58              OR d.id LIKE :likeQ
59              OR d.title LIKE :likeQ
60          ORDER BY score DESC
61          LIMIT 50
62      ";
63  }

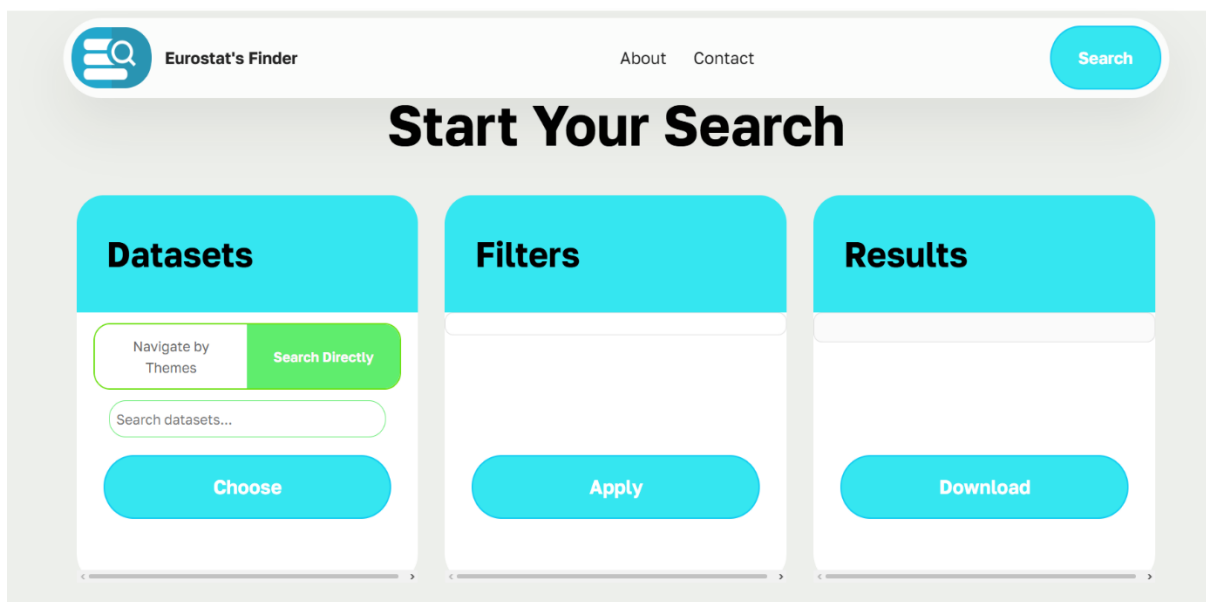
```

Σχήμα 4.7: Υλοποίηση αλγορίθμου κατάταξης και ενίσχυσης πεδίων μέσω σύνθετου SQL ερωτήματος

#### 4.4 Διεπαφή Χρήστη

Εφόσον ολοκληρώθηκε η κατασκευή της Βάσης Δεδομένων, το επόμενο στάδιο είναι η απεικόνιση των πληροφοριών στην διεπαφή χρήστη ώστε να αρχίσει η διαδικασία επιλογής dataset και έπειτα να δοθεί η δυνατότητα δυναμικής επιλογής φίλτρων. Τα φίλτρα εμφανίζονται δυναμικά στην οθόνη μέσω των ασύγχρονων κλήσεων AJAX.

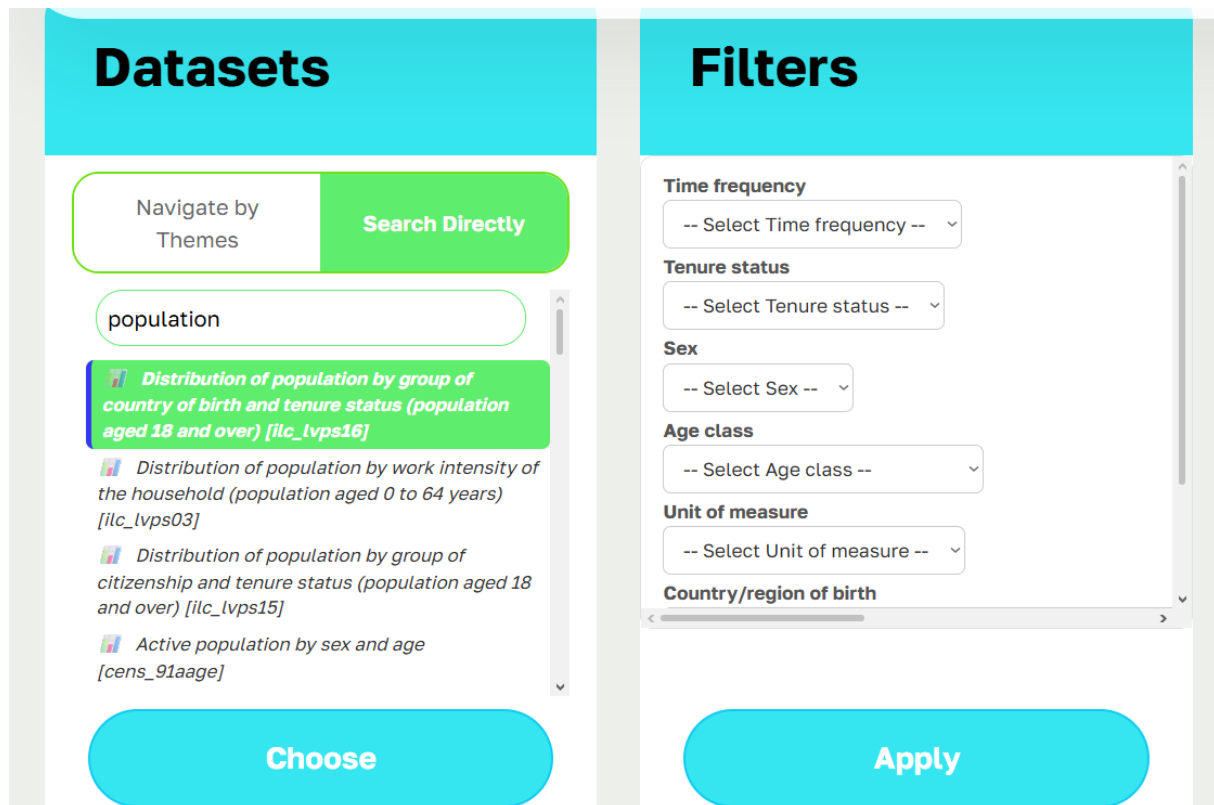
Στο Σχήμα 4.8 απεικονίζεται η αρχική διεπαφή που αντικρίζει ο χρήστης κατά την είσοδό του στην εφαρμογή. Ο σχεδιασμός της βασίζεται σε τρία ξεχωριστά πλαίσια που λειτουργούν ακολουθιακά από τα αριστερά προς τα δεξιά. Πρώτο είναι το πλαίσιο αναζήτησης dataset με την βοήθεια λέξεων-κλειδιών, δεύτερο είναι το πλαίσιο της επιλογής/επιβολής φίλτρων και τρίτο είναι το πλαίσιο που χρησιμοποιείται ως «οθόνη» προβολής των τελικών στατιστικών αποτελεσμάτων.



Σχήμα 4.8: Αρχική Διεπαφή Χρήστη

##### 4.4.1 Δυναμικό Φιλτράρισμα

Μόλις ο χρήστης εισάγει τις λέξεις κλειδιά που επιθυμεί, η εφαρμογή του επιστρέφει τα πενήντα πιο συναφή datasets σε φθίνουσα σειρά. Επιλέγοντας ένα εξ αυτών, αυτομάτως φορτώνονται τα επιμέρους φίλτρα του dataset με drop-down list με κάθε έγκυρη τιμή (Codelists) για την κάθε διάσταση. Στο Σχήμα 4.9, απεικονίζονται οι παραπάνω διαδικασίες καθώς έπειτα από την πληκτρολόγηση του χρήστη population, επιλέχθηκε το πρώτο dataset και εμφανίστηκαν στο διπλανό παράθυρο οι διαστάσεις (dimensions) του με τις πιθανές τιμές τους (Codelists). Στα Codelists εφαρμόστηκε φιλτράρισμα τιμών μέσω του Statistics 1.0 API της Eurostat διότι αρκετά Codelists, όπως αυτό της Γεωγραφίας/GEO, συγκεντρώνει ένα μεγάλο σύνολο τιμών λόγω της αυξημένης συχνότητας που έχει ως διάσταση. Όμως, άλλα datasets αναφέρονται σε πόλεις, άλλα σε χώρες και άλλα σε περιοχές μιας χώρας. Συνεπώς, η ύπαρξη όλων αυτών σε ένα μεγάλο σύνολο θα έδινε ψευδείς επιλογές στον χρήστη που το επιλεγμένο dataset δεν θα υποστήριζε.



Σχήμα 4.9: Τμήμα του UI που δείχνει την εμφάνιση των φίλτρων

#### 4.4.2 Οπτικοποίηση Δεδομένων σε Πίνακα

Αφού ο χρήστης επιλέξει τις τιμές των φίλτρων που θέλει να υποβάλλει, έρχεται το στάδιο της οπτικοποίησης των δεδομένων. Μέσω JavaScript, δημιουργείται ένας πίνακας με αυτόματες περιγραφικές κεφαλίδες και συμπληρωμένα τα κελιά με τις ανακτώμενες τιμές. Εν συνεχεία του προηγούμενου παραδείγματος, στο σχήμα 4.10 απεικονίζεται ο πίνακας οπτικοποίησης καθώς και τα αντίστοιχα φίλτρα που εφαρμόστηκαν. Οι κεφαλίδες Time Period και Value βρίσκονται σε κάθε σετ αποτελεσμάτων ενώ οι υπόλοιπες είναι όσες και τα φίλτρα/dimensions.

The screenshot displays a web interface for data visualization, divided into two main sections: **Filters** and **Results**.

**Filters Section:**

- Tenure status:** Owner [OWN]
- Unit of measure:** Percentage [PC]
- Age class:** -- Select Age class --
- Geopolitical entity (reporting):** Greece [EL]
- Country/region of birth:** Non-EU28 countries (2013-2020) nor reporting country

**Results Section:**

| Time Period | Value | Age class           | Country/region of birth                              | Time freq |
|-------------|-------|---------------------|------------------------------------------------------|-----------|
| 2009        | 32    | From 18 to 54 years | Non-EU28 countries (2013-2020) nor reporting country | Annu      |
| 2010        | 34,6  | From 18 to 54 years | Non-EU28 countries (2013-2020) nor reporting country | Annu      |

At the bottom of each section are buttons labeled **Apply** and **Download**.

Σχήμα 4.10: Οπτικοποίηση δεδομένων σε πίνακα βάσει φίλτρων

#### 4.5 Παραγωγή Σημασιολογικής Δομής

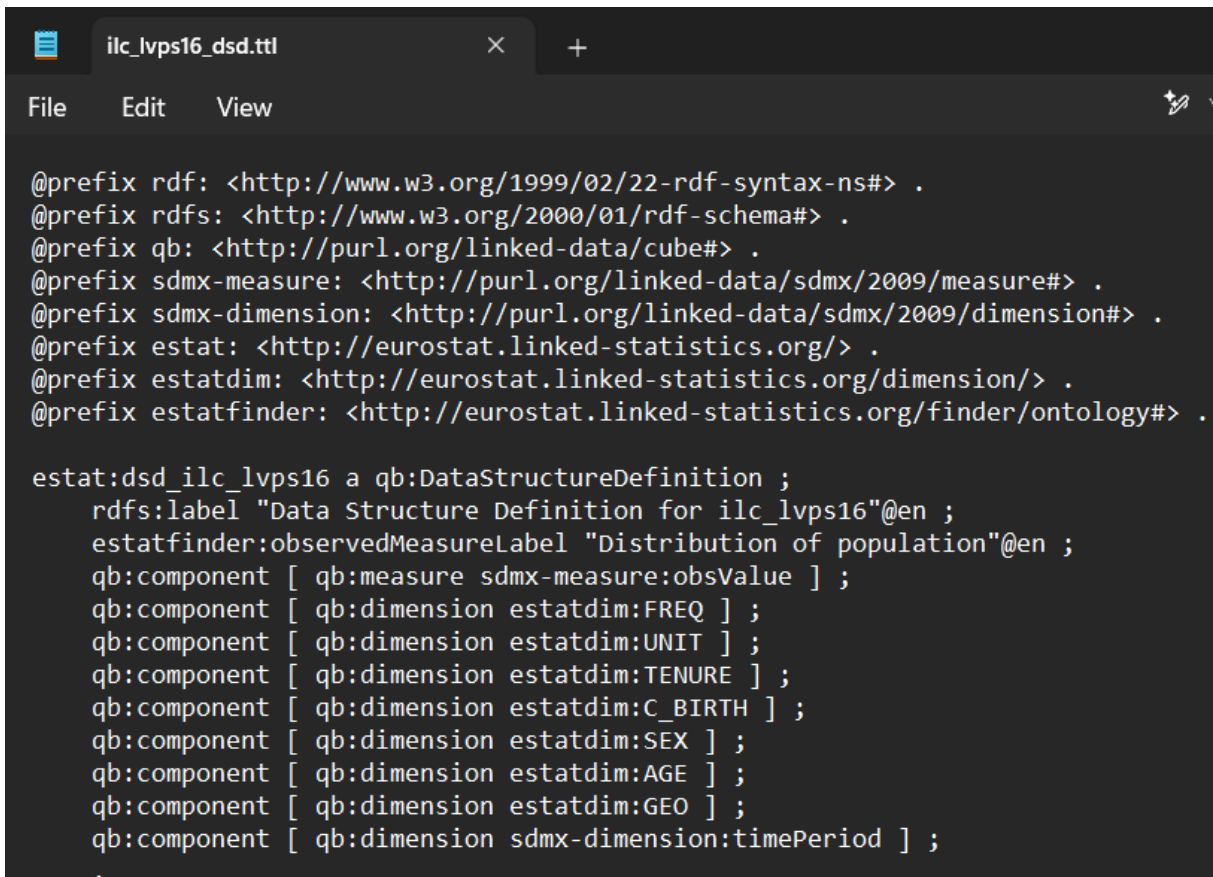
Εκτός από την απλή στατιστική απεικόνιση των αποτελεσμάτων, η εφαρμογή εξάγει τα δεδομένα σε μορφή αναγνώσιμη από μηχανές ώστε να μπορούν να ενταχθούν ομαλά στον Σημασιολογικό Ιστό αξιοποιώντας το επίσημο λεξιλόγιο του W3C, RDF Data Cube και την σύνταξη turtle (.ttl), η οποία είναι εύκολα αναγνώσιμη από ανθρώπους. Μόλις γίνεται η επιλογή dataset, δημιουργείται δυναμικά, και εμφανίζεται στο πλαίσιο των αποτελεσμάτων, ο ορισμός δομής DSD που αποτελεί τον «σκελετό» του κύβου. Συγκεκριμένα, περιέχει, εκτός των απαραίτητων prefixes, το όνομα του dataset και τις διαστάσεις του ονομαστικά, προσφέροντας ένα αφαιρετικό πλάνο για το εκάστοτε dataset. Ο συγκεκριμένος μηχανισμός υλοποιείται μέσω του generate\_dsd.php ενώ στην διεπαφή χρήστη προσφέρεται η δυνατότητα λήψης της δομής DSD σε αρχείο μορφής turtle.

Η ανάγκη αυτή προκύπτει από μία βασική ιδιαιτερότητα της Eurostat: η πλατφόρμα χρησιμοποιεί κοινό δείκτη sdmx-measure:obsValues για όλες τις μετρήσεις, δηλώνοντας ουσιαστικά ότι η τιμή είναι απλώς ένας αριθμός. Καθώς υπάρχει έλλειψη επιμέρους εξειδικευμένων measures για το τι ακριβώς μετράται με αυτόν τον αριθμό, η πληροφορία αυτή βρίσκεται είτε κρυμμένη στον τίτλο του dataset είτε υπάρχει κάποια συγκεκριμένη διάσταση που το ορίζει. Για την επίλυση αυτού του προβλήματος, σχεδιάστηκε μία προσαρμοσμένη (custom) λύση, εισάγοντας νέο πρόθεμα (prefix) για την εφαρμογή: το estatfinder.. Το DSD εμπλουτίζεται με δύο νέες ιδιότητες: το estatfinder:observedMeasureLabel και το estatfinder:measureSelectorDimension.

- A. **estatfinder:observedMeasureLabel:** Εκμεταλλεύεται την πληροφορία που υπάρχει στον τίτλο, απομονώνοντας το μετρήσιμο μέγεθος με την αποκοπή περιττού κειμένου, πρωτίστως

βάσει της ύπαρξης παύλας και δευτερευόντως της λέξης by. Παραδείγματος χάρη στο dataset με τίτλο: «Distribution of population by sex and age», απομονώνεται ως label το «distribution of population».

- B. **estatfinder:measureSelectorDimension**: Σαρώνει τις διαστάσεις του εκάστοτε dataset ψάχνοντας γνωστές λέξεις διαστάσεων που φανερώνουν ότι είναι δείκτες (π.χ. na\_item, indic, post, int\_rt). Σε περίπτωση εύρεσης ενός, δηλώνεται ρητά στον κύβο ποια διάσταση έχει τον ρόλο του επιλογέα μεγέθους μέτρησης ενώ σε περίπτωση μη εντοπισμού, δεν εμφανίζεται καθόλου.



```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix qb: <http://purl.org/linked-data/cube#> .
@prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#> .
@prefix sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#> .
@prefix estat: <http://eurostat.linked-statistics.org/> .
@prefix estatdim: <http://eurostat.linked-statistics.org/dimension/> .
@prefix estatfinder: <http://eurostat.linked-statistics.org/finder/ontology#> .

estat:dsd_ilc_lvps16 a qb:DataStructureDefinition ;
  rdfs:label "Data Structure Definition for ilc_lvps16"@en ;
  estatfinder:observedMeasureLabel "Distribution of population"@en ;
  qb:component [ qb:measure sdmx-measure:obsValue ] ;
  qb:component [ qb:dimension estatdim:FREQ ] ;
  qb:component [ qb:dimension estatdim:UNIT ] ;
  qb:component [ qb:dimension estatdim:TENURE ] ;
  qb:component [ qb:dimension estatdim:C_BIRTH ] ;
  qb:component [ qb:dimension estatdim:SEX ] ;
  qb:component [ qb:dimension estatdim:AGE ] ;
  qb:component [ qb:dimension estatdim:GEO ] ;
  qb:component [ qb:dimension sdmx-dimension:timePeriod ] ;
  .
```

Σχήμα 4.11: DSD δομή σε μορφή .ttl

Μετά και την υποβολή των φίλτρων, το αρχείο convert.php είναι υπεύθυνο παραγωγής του νέου RDF αρχείο το οποίο σε αντίθεση με το abstract DSD, περιέχει πολλές πληροφορίες για το dataset. Στο παραγόμενο αρχείο turtle, ορίζεται η ταυτότητα qb:Dataset και χρησιμοποιούνται βασικά στοιχεία από τα πρότυπα DCAT και DCTerms. Συγκεκριμένα, ο τίτλος και η εκτενής περιγραφή αποδίδονται με τα κατηγορήματα dcterms:title και dcterms:description αντίστοιχα ενώ το θέμα μέσω του dcat:theme. Όσο αναφορά τις σημασιολογικές ετικέτες EuroVoc που παράχθηκαν στο στάδιο της ευρετηρίασης, ορίζονται στο γράφημα ως dcat:keyword εμπλουτίζοντας την σημασιολογική και θεματική διασύνδεση. Στην συνέχεια, κάθε στατιστική τιμή μοντελοποιείται ξεχωριστά ως qb:Observation και του αποδίδονται οι κατάλληλες διαστάσεις τύπου estatdim:{διάσταση} και η μετρήσιμη τιμή ως sdmx-measure:obsValue {τιμή} συνοδευόμενο από αντιπροσωπευτικό όρο της μετρήσιμης τιμής (π.χ. decimal).

Τέλος, εμφανίζονται τα Codelabels που περιλαμβάνει όλους τους μοναδικούς κωδικούς που εντοπίστηκαν στις τιμές των αντίστοιχων dimensions των παρατηρήσεων (obsValues) και τις επεξηγεί χρησιμοποιώντας το λεξιλόγιο SKOS. Κάθε κωδικός ορίζεται από μία έννοια ως skos:Concept και του αποδίδει ως skos:notation την κωδικοποιημένη τιμή της διάστασης (π.χ. EL κωδικός του Greece ), ως rdfs:label και skos:prefLabel την αντίστοιχη ανθρώπινα αναγνώσιμη τιμή του. Αυτός ο μηχανισμός διασφαλίζει την αυτοτέλεια του παραγόμενου αρχείου, ενισχύοντας την διαλειτουργικότητα της εφαρμογής.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix qb: <http://purl.org/linked-data/cube#> .
@prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#> .
@prefix sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#> .
@prefix sdmx-attribute: <http://purl.org/linked-data/sdmx/2009/attribute#> .
@prefix estat: <http://eurostat.linked-statistics.org/> .
@prefix estatdim: <http://eurostat.linked-statistics.org/dimension/> .
@prefix estatcode: <http://eurostat.linked-statistics.org/code/> .

# Dataset

estat:dataset_nama_10_gdp a qb:DataSet ;
    dcterms:identifier "nama_10_gdp" ;
    dcterms:title "Gross domestic product (GDP) and main components (output, expenditure and income) - annual data"@en ;
    rdfs:label "Gross domestic product (GDP) and main components (output, expenditure and income) - annual data"@en ;
    dcterms:description "National accounts are a coherent and consistent set of macroeconomic indicators, which provide an overall picture of the economic situation and are widely used for economic analysis and forecasting, policy design and policy making. Annual national accounts are compiled in accordance with the European System of Accounts - ESA 2010 as defined in Annex B of the Council Regulation (EU) No 549/2013 of the European Parliament and of the Council of 21 May 2013, amended by Council Regulation (EU) 2023/734 of 15 March 2023. Gross domestic product (GDP) is one of the key aggregates in the European system of accounts (ESA). GDP is a measure of the total economic activity taking place on an economic territory which leads to
```

Σχήμα 4.12: Δήλωση prefixes, τίτλου και περιγραφής σε παραγόμενο αρχείο RDF με παρατηρήσεις

```

dcats:theme "General and regional statistics"@en ;
dcats:keyword "non-profit institutions serving households"@en ;
dcats:keyword "government final consumption expenditure"@en ;
dcats:keyword "european system of accounts 2010"@en ;
dcats:keyword "general and regional statistics"@en ;
dcats:keyword "gross fixed capital formation"@en ;
dcats:keyword "european system of accounts"@en ;
dcats:keyword "intermediate consumption"@en ;
dcats:keyword "consumption expenditure"@en ;
dcats:keyword "gross domestic product"@en ;
dcats:keyword "candidate countries"@en ;
dcats:keyword "economic indicators"@en ;
dcats:keyword "european parliament"@en ;
dcats:keyword "regional statistics"@en ;
dcats:keyword "council regulation"@en ;
dcats:keyword "economic indicator"@en ;
qb:structure estat:dsd_nama_10_gdp .

# Observations

estat:obs_nama_10_gdp_FREQ-A_GEO-EL_NA_ITEM-B1G_UNIT-CLV_I20_time-1995 a qb:Observation ;
qb:dataSet estat:dataset_nama_10_gdp ;
sdmx-dimension:timePeriod "1995"^^xsd:gYear ;
sdmx-measure:obsValue "89.114"^^xsd:decimal ;
estatdim:FREQ estatcode:FREQ_A ;
estatdim:GEO estatcode:GEO_EL ;
estatdim:NA_ITEM estatcode:NA_ITEM_B1G ;

```

Σχήμα 4.13: Δήλωση ετικετών EuroVoc και αρχή των παρατηρήσεων σε αρχείο RDF με παρατηρήσεις

```

# Code Labels

estatcode:FREQ_A a skos:Concept ;
  skos:notation "A" ;
  rdfs:label "Annual"@en ;
  skos:prefLabel "Annual"@en .

estatcode:GEO_EL a skos:Concept ;
  skos:notation "EL" ;
  rdfs:label "Greece"@en ;
  skos:prefLabel "Greece"@en .

estatcode:NA_ITEM_B1G a skos:Concept ;
  skos:notation "B1G" ;
  rdfs:label "Value added, gross"@en ;
  skos:prefLabel "Value added, gross"@en .

estatcode:UNIT_CLV_I20 a skos:Concept ;
  skos:notation "CLV_I20" ;
  rdfs:label "Chain linked volumes, index 2020=100"@en ;
  skos:prefLabel "Chain linked volumes, index 2020=100"@en .

```

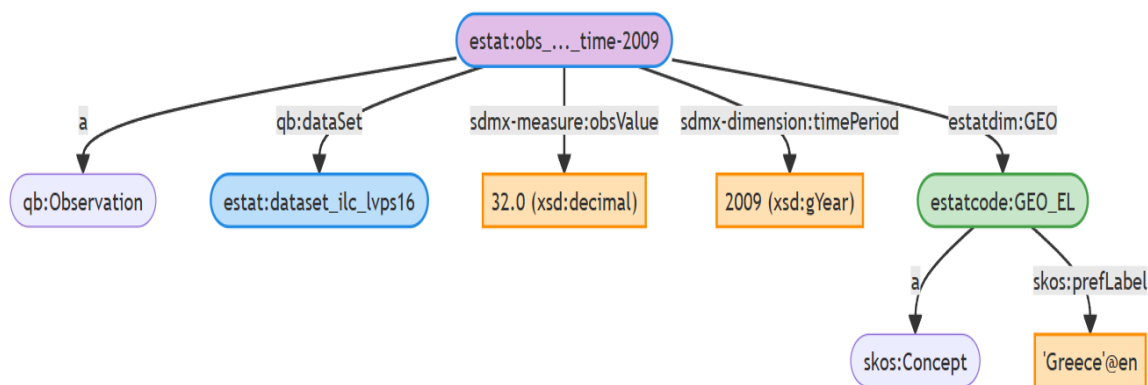
Σχήμα 4.14: Δήλωση Codelabels σε παραγόμενο RDF με παρατηρήσεις

Για την καλύτερη κατανόηση της σημασιολογικής μοντελοποίησης που εφαρμόζεται στο σύστημα, ακολουθεί ο Πίνακας 4.1 που παραθέτει την άμεση αντιστοίχιση των πεδίων της Eurostat με τους επίσημους όρους και ιδιότητες καθιερωμένων λεξιλογίων του Σημασιολογικού Ιστού (RDF Data Cube, SKOS, DCAT, Dublin Core).

Πίνακας 4.1: Αντιστοίχιση των πεδίων της Eurostat με επίσημους όρους σημασιολογικής ιδιότητας

| Πεδίο Eurostat    | Σημασιολογική Ιδιότητα |
|-------------------|------------------------|
| Αναγνωριστικό ID  | dcterms:identifier     |
| Τίτλος            | dcterms:title          |
| Περιγραφή         | dcterms:description    |
| Θέμα              | dcat:theme             |
| Ετικέτες EuroVoc  | dcat:keyword           |
| Μετρήσιμη Τιμή    | sdmx-measure:obsValue  |
| Γραμμή Δεδομένων  | qb:Observation         |
| Κωδικός Διάστασης | skos:Concept           |

Τέλος, ακολουθεί το Σχήμα 4.15 για να συνοψίσει και να αναδείξει οπτικά την δομή του RDF Data Cube, υπενθυμίζοντας ότι τα στατιστικά δεδομένα δεν είναι απλές γραμμές ενός πίνακα αλλά αποτελούν κόμβους ενός διασυνδεδεμένου γραφήματος με σημασιολογικές έννοιες.



Σχήμα 4.15: Διάγραμμα RDF Data Cube ενός τυχαίου παραδείγματος σε τυχαίο dataset

#### 4.6 Μηχανισμός Διαμεσολάβησης και Διαχείρισης CORS

Η ανάπτυξη μιας διαδικτυακής εφαρμογής που αλληλεπιδρά συνεχώς με εξωτερικά APIs, έχει να αντιμετωπίσει τους περιορισμούς ασφαλείας των σύγχρονων φυλλομετρητών και συγκεκριμένα την πολιτική Cross-Origin Resource Sharing (CORS). Αν και επινοήθηκε για την προστασία των χρηστών από κακόβουλες επιθέσεις, όπως το Cross-Site Request Forgery, η διαμόρφωσή του συχνά αποτελεί εμπόδιο και ευθύνεται για ευπάθειες και υποβάθμιση της απόδοσης των διαδικτυακών εφαρμογών [25].

Βάσει αυτής, εάν η διεπαφή χρήστη επιχειρήσει να εκτελέσει απευθείας ασύγχρονες κλήσεις AJAX προς τους διακομιστές της Eurostat, το αίτημα θα απορριφθεί για λόγους ασφαλείας. Η πολιτική CORS απορρίπτει ό,τι δεν έχει ρητή άδεια πρόσβασης από τους ίδιους τους διακομιστές σε επίπεδο διεπαφής χρήστη [26].

Όμως δεν ισχύει το ίδιο για την επικοινωνία μεταξύ διακομιστών. Με τη δημιουργία ενός αρχείου PHP ξεπερνιέται το πρόβλημα της πολιτικής CORS, εφόσον το αρχείο PHP λειτουργεί ως διαμεσολαβητής (proxy) μεταξύ του κώδικα JS του αρχείου διεπαφής χρήστη και των διακομιστών της Eurostat, αναλαμβάνοντας την απευθείας επικοινωνία server-to-server με τα APIs. Στην επικοινωνία server-to-server δεν υπάρχει κανένας περιορισμός οπότε οι κλήσεις επιστρέφουν σωστά τα δεδομένα.

Η παραπάνω προσέγγιση πέραν της λύσης στους περιορισμούς ασφαλείας CORS, επιτρέπει στην εφαρμογή να διαχειρίζεται καλύτερα τα σφάλματα HTTP, δίνοντας αποδοτικότερα μηνύματα ανατροφοδότησης στους χρήστες όταν συμβαίνει κάποιο σφάλμα.

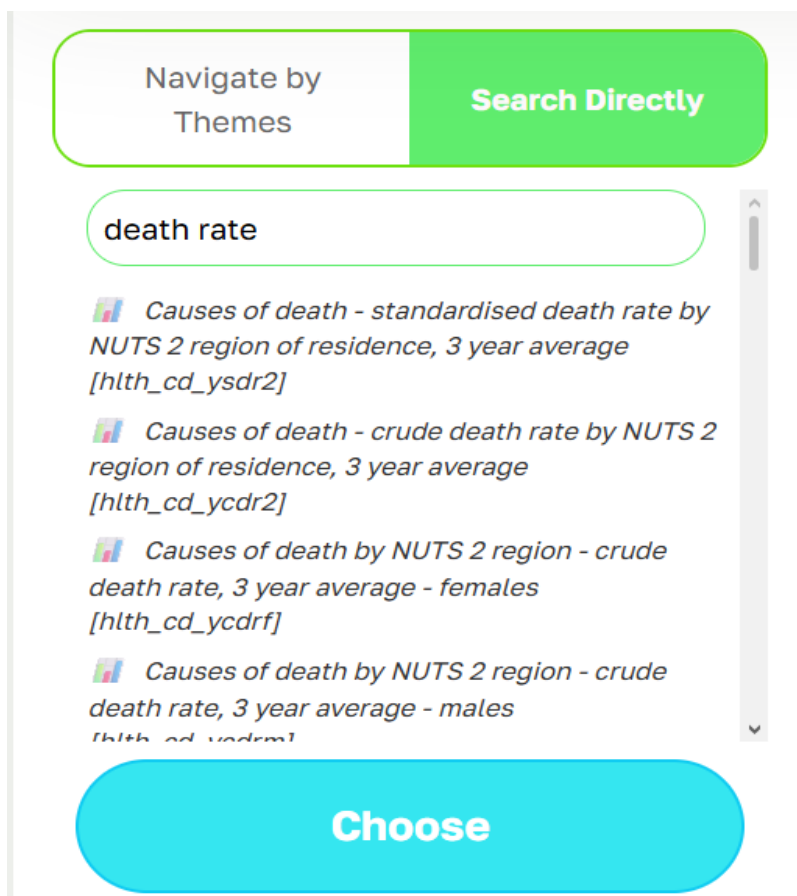
## Κεφάλαιο 5ο: Αποτελέσματα και Αξιολόγηση

### 5.1 Σενάρια Χρήσης

Ο τεράστιος όγκος δεδομένων της Eurostat, χρήζει της ύπαρξης πολλαπλών σεναρίων χρήσης ώστε να αξιολογηθεί σωστά η μηχανή αναζήτησης αλλά και να δικαιολογηθεί η κατανομή βαρών στα πεδία. Η μηχανή αναζήτησης θέτει ως σκοπό την ταύτιση, όχι μόνο από τον τίτλο και το αναγνωριστικό id των datasets, αλλά και μέσω του σημασιολογικού εμπλουτισμού που πραγματοποιήθηκε. Για την καλύτερη κατανόηση του μηχανισμού της μηχανής αναζήτησης, τα πολλαπλά σενάρια χρήσεις μπορούν να ομαδοποιηθούν σε δύο μεγάλες κατηγορίες: σενάριο άμεσης λεκτικής ταύτισης με τίτλο ή αναγνωριστικό ID και το σενάριο ταύτισης με όρους σημασιολογικής αναζήτησης.

#### 5.1.1 Σενάριο Άμεσης Λεκτικής Ταύτισης

Σε αυτό το σενάριο, ελέγχεται η συμπεριφορά της εφαρμογής σε ερωτήματα χρηστών που γνωρίζουν τι ψάχνουν και εισάγουν κάποιους όρους του τίτλου ή το αναγνωριστικό ID. Ενώ η εισαγωγή ενός όρου του τίτλου μόνο δεν εγγυάται την ύπαρξη του επιθυμητού dataset στην κορυφή της λίστας των υποψηφίων, η εισαγωγή από δύο ταυτόσημους όρους και πάνω με τον τίτλο πρέπει να επιστρέφει στην κορυφή της λίστας το επιθυμητό dataset. Όπως διατυπώθηκε στην αρχιτεκτονική της βάσης, το ταίριασμα στον τίτλο λαμβάνει το μεγαλύτερο βάρος ίσο με 100, ενώ η ταύτιση του ID λαμβάνει το δεύτερο μεγαλύτερο βάρος ίσο με 40. Η εισαγωγή ολόκληρου του ID επιστρέφει στην κορυφή της λίστας το αντίστοιχο dataset.



The screenshot shows a web interface for searching datasets. At the top, there are two buttons: "Navigate by Themes" (white with a green border) and "Search Directly" (solid green). Below these is a search bar containing the text "death rate". To the right of the search bar is a vertical scrollbar. Below the search bar, there is a list of four search results, each preceded by a small bar chart icon. The results are:

- Causes of death - standardised death rate by NUTS 2 region of residence, 3 year average*  
[hlth\_cd\_ysdr2]
- Causes of death - crude death rate by NUTS 2 region of residence, 3 year average*  
[hlth\_cd\_ycdr2]
- Causes of death by NUTS 2 region - crude death rate, 3 year average - females*  
[hlth\_cd\_ycdrf]
- Causes of death by NUTS 2 region - crude death rate, 3 year average - males*  
[hlth\_cd\_ycdrm]

At the bottom of the interface is a large, rounded blue button labeled "Choose".

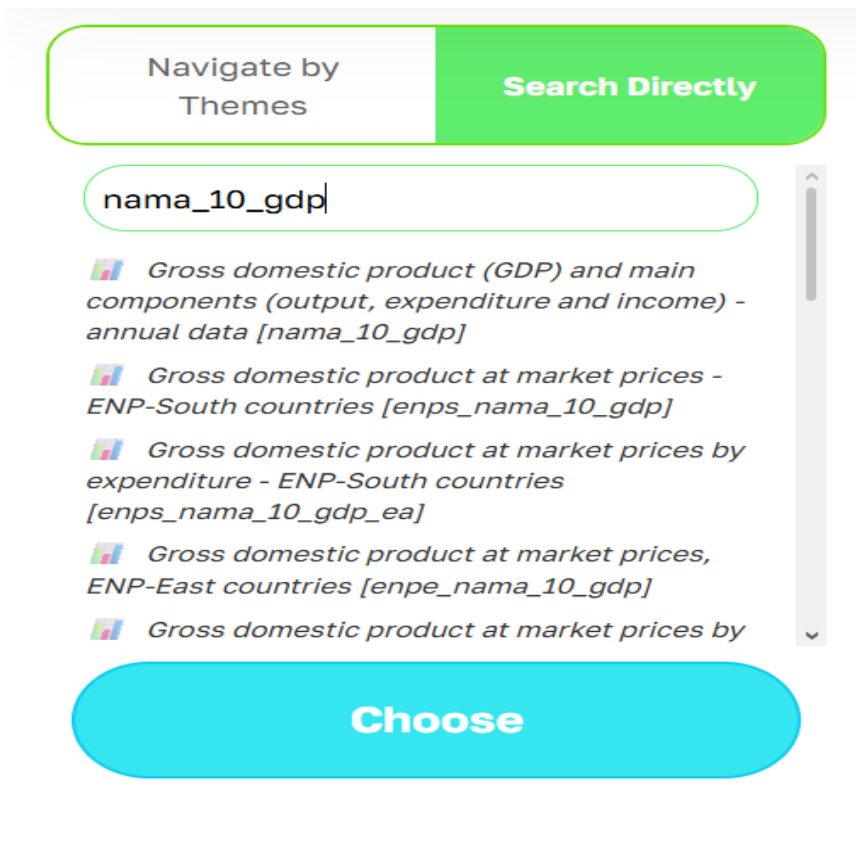
Σχήμα 5.1: Σενάριο άμεσης λεκτικής ταύτισης με τίτλο

Ομορφή Εκτύπωση ☒

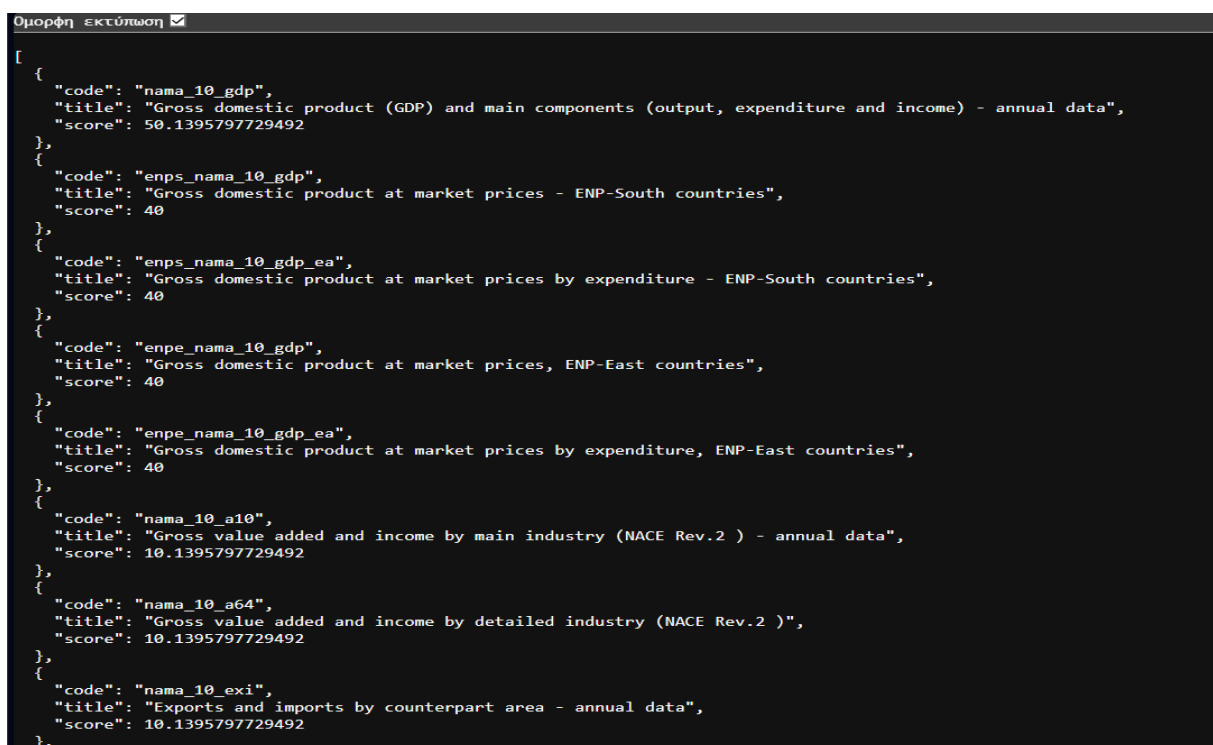
```
[
  {
    "code": "hlth_cd_ysdr2",
    "title": "Causes of death - standardised death rate by NUTS 2 region of residence, 3 year average",
    "score": 1075.37178039551
  },
  {
    "code": "hlth_cd_ycdr2",
    "title": "Causes of death - crude death rate by NUTS 2 region of residence, 3 year average",
    "score": 1075.37178039551
  },
  {
    "code": "hlth_cd_ycdrf",
    "title": "Causes of death by NUTS 2 region - crude death rate, 3 year average - females",
    "score": 1075.37178039551
  },
  {
    "code": "hlth_cd_ycdrm",
    "title": "Causes of death by NUTS 2 region - crude death rate, 3 year average - males",
    "score": 1075.37178039551
  },
  {
    "code": "hlth_cd_ycdrt",
    "title": "Causes of death by NUTS 2 region - crude death rate, 3 year average - total",
    "score": 1075.37178039551
  },
  {
    "code": "hlth_cd_msdr2",
    "title": "Causes of death - standardised death rate by country of residence, monthly data",
    "score": 1074.58805084229
  },
  {
    "code": "hlth_cd_ysdr1",
    "title": "Causes of death by NUTS 2 region - standardised death rate, 3 year average",
    "score": 1074.58805084229
  },
  {
    "code": "hlth_cd_acdr2",
    "title": "Causes of death - crude death rate by NUTS 2 region of residence",
    "score": 1070.14221191406
  }
]
```

Σχήμα 5.2: Εμφάνιση score του παραπάνω παραδείγματος

Τα Σχήματα 5.1 και 5.2, τονίζουν την δυναμική που υπάρχει στην ταύτιση όρων εισαγωγής με λέξεις τίτλου. Συγκεκριμένα, τα πρώτα αποτελέσματα περιέχουν και τους δύο όρους αναζήτησης με αποτέλεσμα να εισπράττουν το μεγαλύτερο βάρος από τον μηχανισμό. Είναι εμφανές ότι η αναζήτηση ταυτίζεται εξ ολοκλήρου με αρκετά datasets που παρουσιάζουν υψηλή θεματική συνάφεια αλλά διαφέρουν στις διαστάσεις (π.χ. γεωγραφικά). Οι βαθμολογίες άνω των χιλίων βαθμών επιβεβαιώνει τα παραπάνω, οι όροι αναζήτησης βρέθηκαν πολλές φορές και σε άλλα πεδία βαθμολόγησης.



Σχήμα 5.3: Σενάριο εισαγωγής αναγνωριστικού ID



Σχήμα 5.4: Εμφάνιση score του παραπάνω παραδείγματος

Τα Σχήματα 5.3 και 5.4, αναδεικνύουν την άρρηκτη σχέση μεταξύ αναγνωριστικού ID και αναζήτησης. Τα αναγνωριστικά συχνά μοιάζουν μεταξύ τους αν αναφέρονται σε ίδιο θέμα και σε περιπτώσεις, το ID κάποιου γενικού dataset εμπεριέχεται ολόκληρο σε άλλο ID ειδικότερου dataset (π.χ. nama\_10\_gdp και enps\_nama\_10\_gdp). Στο παράδειγμα, ο μηχανισμός αναζήτησης δίνει προβάδισμα στο dataset με ακριβώς αυτό το αναγνωριστικό που εισήχθη και έπονται τα υπόλοιπα datasets που το ID τους εμπεριέχει το εισαγόμενο κείμενο αναζήτησης. Ειδικότερα από το Σχήμα 5.4, συμπεραίνεται η κλίμακα τάυτισης βάσει του score. Το επιθυμητό dataset με id:nama\_10\_gdp λαμβάνει 40 βαθμούς από το ταίριασμα αναγνωριστικού και τους υπόλοιπους βαθμούς από τον σημασιολογικό εμπλουτισμό μεταδεδομένων ESMS. Αντιθέτως, τα υπόλοιπα datasets που το εμπεριέχουν ολόκληρο μες το δικό τους αναγνωριστικό ID λαμβάνουν ακριβώς 40 βαθμούς ενώ τέλος τα υπόλοιπα datasets επιβραβεύονται με βαθμούς μόνο από τα μεταδεδομένα ESMS.

### 5.1.2 Σενάριο Ταύτισης μέσω Σημασιολογικής Αναζήτησης

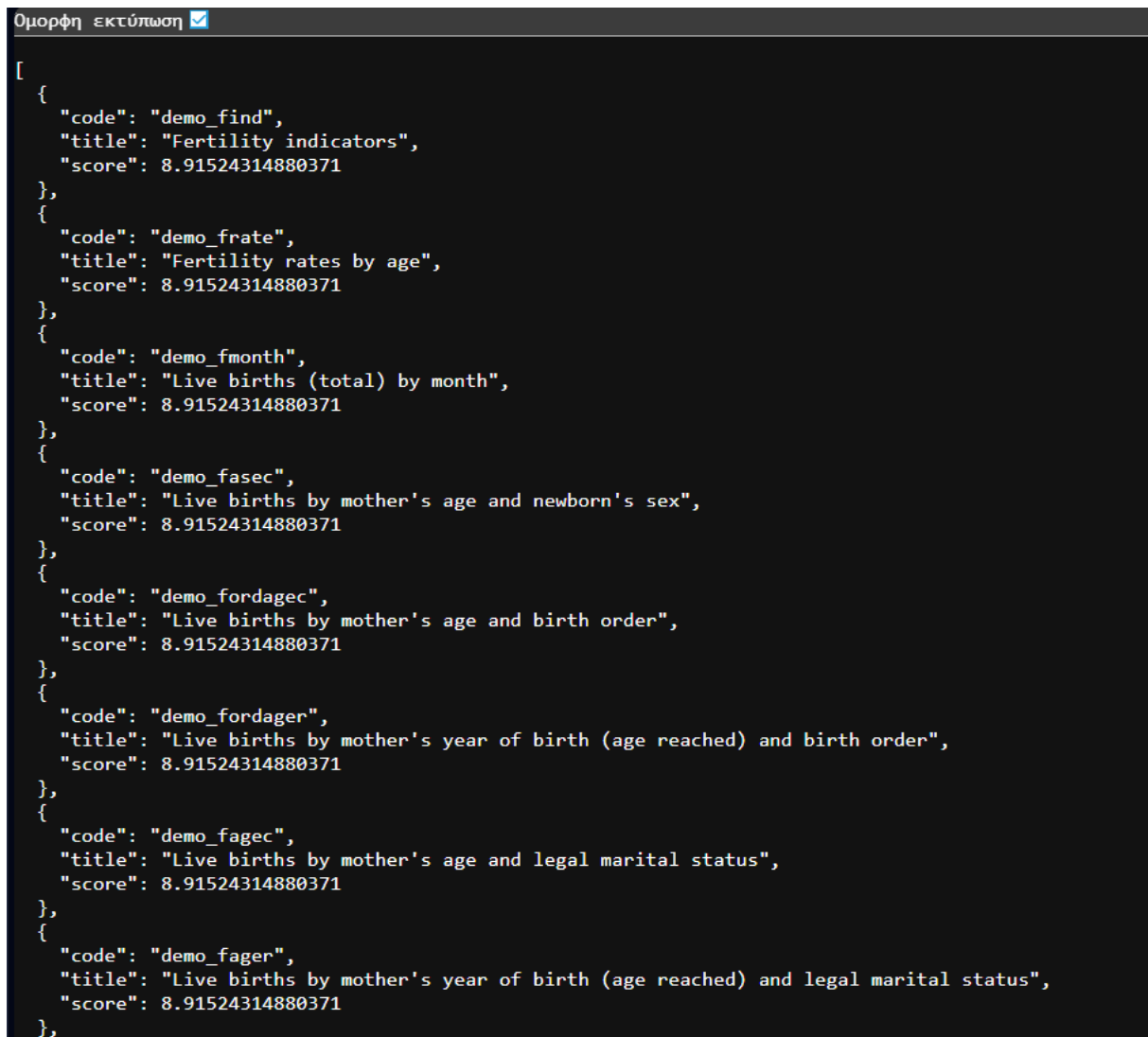
Το δεύτερο σενάριο το οποίο έρχεται σε αντιδιαστολή με τις συμβατικές μηχανές αναζήτησης της Eurostat που καλύπτουν αυστηρά μόνο το πρώτο σενάριο της άμεσης ταύτισης με τον τίτλο. Η αναζήτηση στην εφαρμογή λειτουργεί σε βάθος, αξιοποιώντας τα πεδία του EuroVoc με την χρήση της εντολής LIKE και βάρος ίσο με 15 και το Full-Text Indexing των μεταδεδομένων ESMS με βάρος ίσο με 10. Με αυτόν τον τρόπο ανακτώνται «κρυμμένα» datasets των οποίων ο τίτλος μπορεί να διαφέρει από τους εισαγόμενους όρους αλλά σημασιολογικά ταυτίζονται με τα ζητούμενα του χρήστη.

The screenshot shows a search interface with two main buttons at the top: "Navigate by Themes" (light green) and "Search Directly" (dark green). Below these is a search input field containing the text "pregnancy". To the right of the input field is a vertical scrollbar. Below the input field is a list of search results, each preceded by a small icon and followed by a dataset ID in brackets. The results are:

- Fertility indicators [demo\_find]
- Fertility rates by age [demo\_frate]
- Live births (total) by month [demo\_fmonth]
- Live births by mother's age and newborn's sex [demo\_fasec]
- Live births by mother's age and birth order [demo\_fordagec]
- Live births by mother's year of birth (age reached) and birth order [demo\_fordager]
- Live births by mother's age and legal marital

At the bottom of the interface is a large blue button labeled "Choose".

Σχήμα 5.5: Σενάριο εισαγωγής όρου εκτός τίτλου



```

Ομορφη ΕΚΤΥΠΩΣΗ ☒
[
  {
    "code": "demo_find",
    "title": "Fertility indicators",
    "score": 8.91524314880371
  },
  {
    "code": "demo_frate",
    "title": "Fertility rates by age",
    "score": 8.91524314880371
  },
  {
    "code": "demo_fmonth",
    "title": "Live births (total) by month",
    "score": 8.91524314880371
  },
  {
    "code": "demo_fasec",
    "title": "Live births by mother's age and newborn's sex",
    "score": 8.91524314880371
  },
  {
    "code": "demo_fordagec",
    "title": "Live births by mother's age and birth order",
    "score": 8.91524314880371
  },
  {
    "code": "demo_fordager",
    "title": "Live births by mother's year of birth (age reached) and birth order",
    "score": 8.91524314880371
  },
  {
    "code": "demo_fagec",
    "title": "Live births by mother's age and legal marital status",
    "score": 8.91524314880371
  },
  {
    "code": "demo_fager",
    "title": "Live births by mother's year of birth (age reached) and legal marital status",
    "score": 8.91524314880371
  },
]

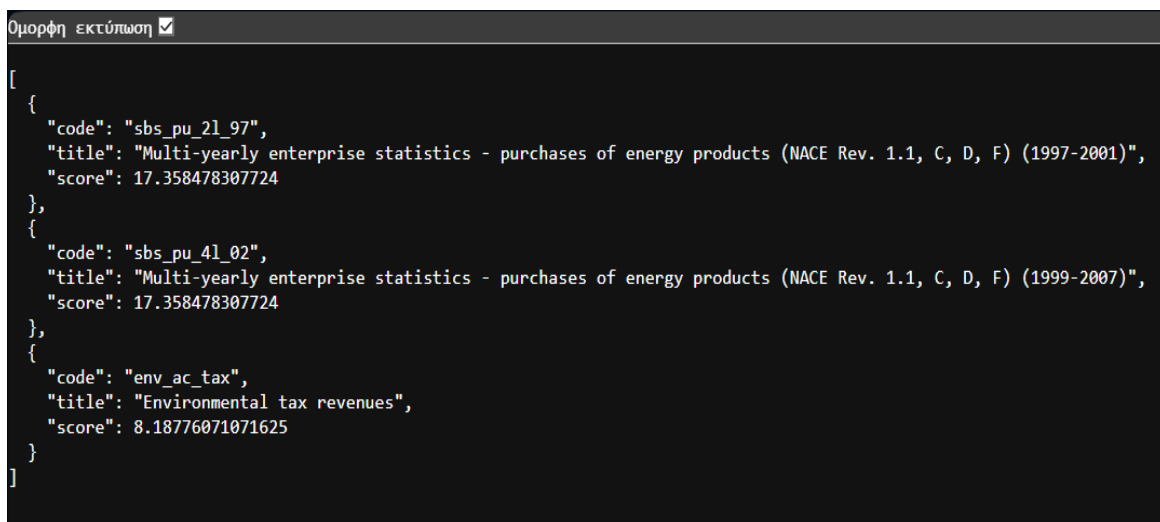
```

Σχήμα 5.6: Εμφάνιση score του παραπάνω παραδείγματος

Βάσει των Σχημάτων 5.5 και 5.6, το σενάριο της εισαγωγής του όρου «pregnancy» δεν ταυτίζεται με κάποιον τίτλο ή ID. Όμως βάσει του σημασιολογικού εμπλουτισμού με τις ετικέτες της Eurovoc και με την αξιοποίηση των μεταδεδομένων ESMS, επιστρέφονται στον χρήστη datasets που συνδέονται σημασιολογικά με την εγκυμοσύνη αφού τα πρώτα στην κατάταξη αναφέρουν δείκτες γονιμότητας και στατιστικά γεννών. Οι χαμηλές βαθμολογίες κάτω του 10, υποδεικνύουν ότι υπήρξε ταίριασμα στα εκτεταμένα κείμενα ESMS ενώ η συγκομιδή ίδιου score σημαίνει ότι τα datasets περιείχαν τον όρο αναζήτησης ίσες φορές, χωρίς κάποιο να ξεχωρίζει.

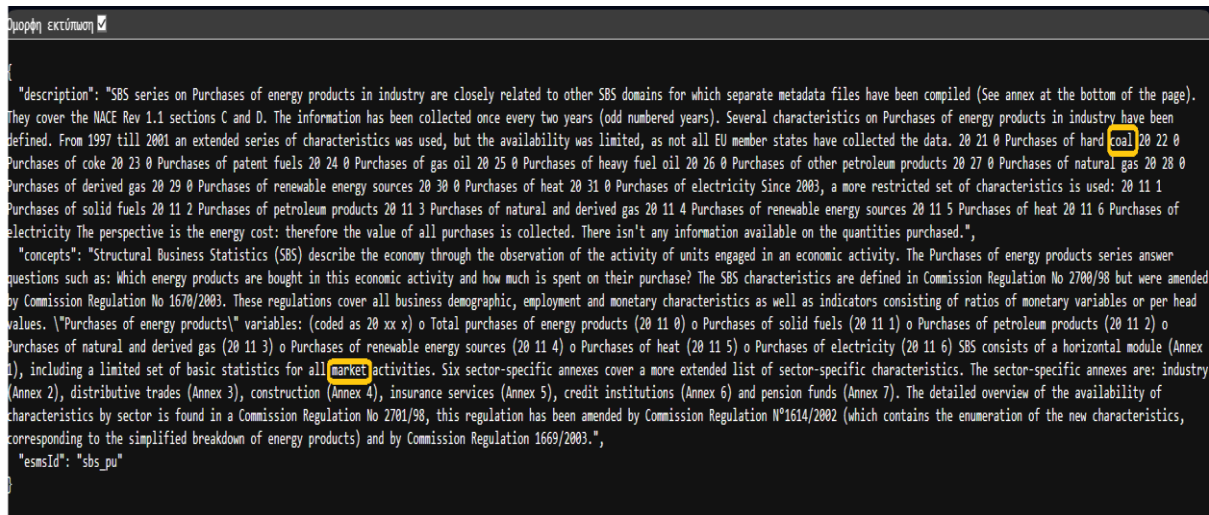


Σχήμα 5.7: Σενάριο εισαγωγής δύο λέξεων εκτός τίτλου στην μηχανή αναζήτησης



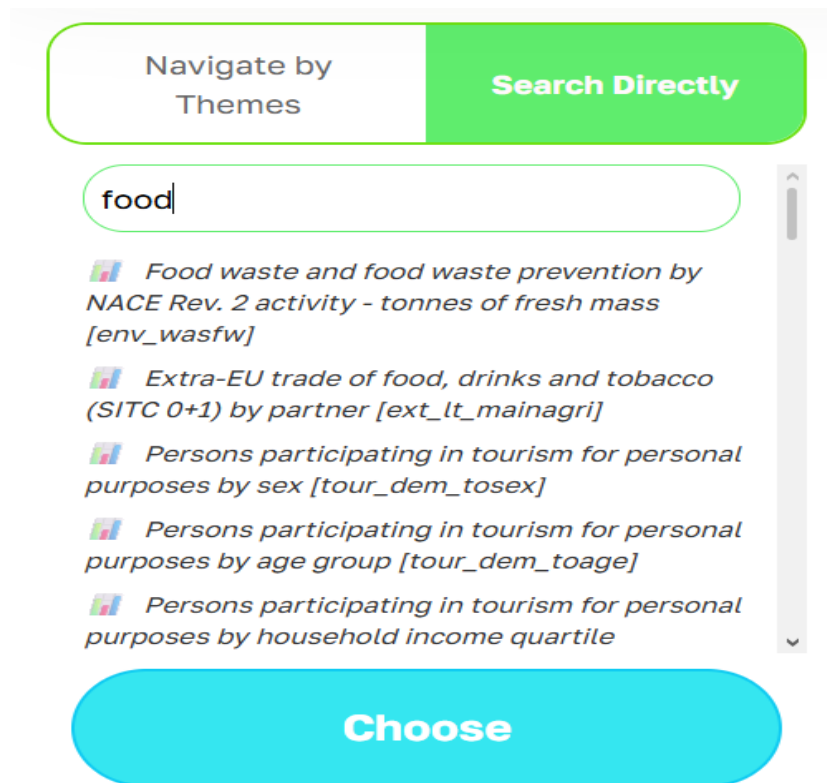
Σχήμα 5.8: Εμφάνιση score του παραπάνω παραδείγματος

Στο παράδειγμα των σχημάτων 5.7 και 5.8 βλέπουμε πως ανταποκρίνεται η μηχανή αναζήτησης σε εισαγωγή δύο όρων εκτός τίτλου. Τα αποτελέσματα είναι μόνο τρία με σχετικά χαμηλή βαθμολογία. Αυτό συμβαίνει λόγω της έλλειψης ταιριάσματος σε τίτλο και ID. Έστω και ένας όρος να υπήρχε στον τίτλο κάποιων datasets, θα ανέβαιναν αυτά στην κορυφή της κατάταξης εφόσον οι υπόλοιποι όροι συγκέντρωναν μεμονωμένα μη μηδενική βαθμολογία. Στην συγκεκριμένη περίπτωση τα δύο πρώτα datasets συγκεντρώνουν την ίδια βαθμολογία καθώς μοιράζονται το ίδιο αρχείο ESMS sbs\_pu, το μοναδικό ταίριασμα που επιτυγχάνουν βρίσκεται σε αυτό το αρχείο. Στο παρακάτω σχήμα απεικονίζεται το παραπάνω ESMS με κυκλωμένα με κίτρινο χρώμα, τα ταιριάσματα βάσει των όρων αναζήτησης «coal market».

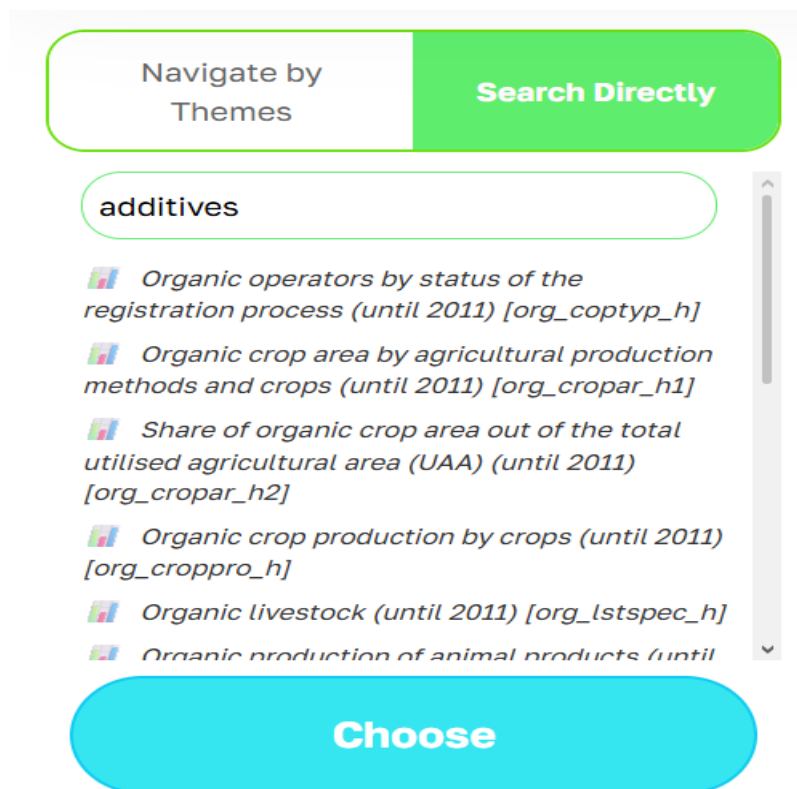


Σχήμα 5.9: Αρχείο ESMS sbs\_pu

Όμως όπως αναφέρθηκε παραπάνω, στην περίπτωση εισαγωγής από δύο και πάνω όρους, η αναζήτηση λειτουργεί ξεχωριστά για τον καθένα. Αν κάποιος από αυτούς, συγκεντρώσει μηδενική βαθμολογία σε dataset τότε αυτό αποκλείεται από τα datasets προς εμφάνιση, λαμβάνοντας μηδενική βαθμολογία. Ο παραπάνω μηχανισμός αυστηρού φιλτραρίσματος ονομάζεται Cross-Field AND Matching. Στα παρακάτω σχήματα, απεικονίζεται ένα παράδειγμα ίδιας νοοτροπίας με δύο εισαγόμενους όρους. Ο καθένας συγκεντρώνει ξεχωριστά σε αναζήτηση πλήθος συναφών αποτελεσμάτων αλλά με την ταυτόχρονη εισαγωγή τους, δεν εμφανίζεται κανένα αποτέλεσμα.



Σχήμα 5.10: Εισαγωγή του όρου «food»

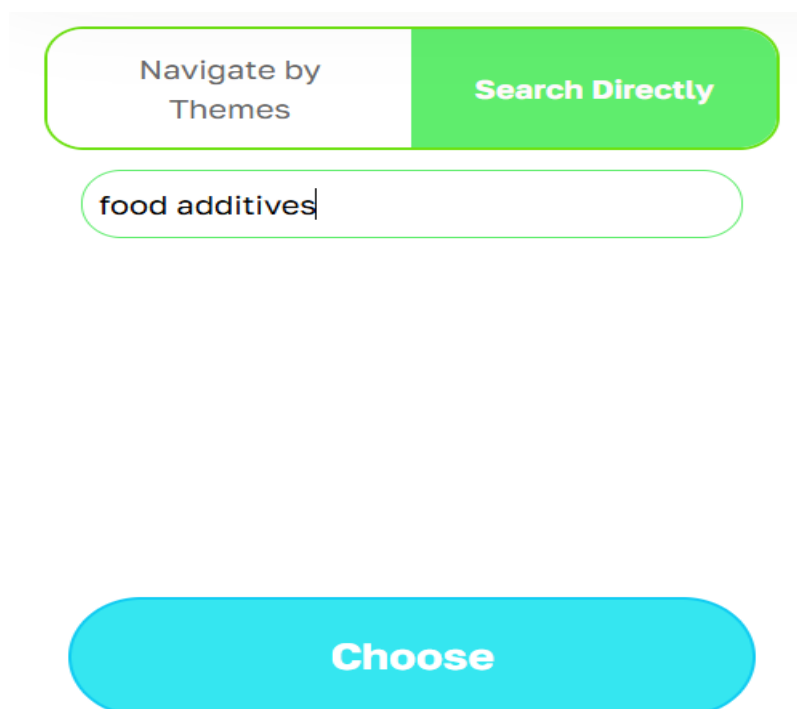


The interface features a top navigation bar with two buttons: "Navigate by Themes" and "Search Directly". Below this is a search input field containing the text "additives". To the right of the input field is a vertical scrollbar. Below the input field, a list of search results is displayed, each preceded by a small icon and followed by a code in brackets. The results are:

- Organic operators by status of the registration process (until 2011) [org\_coptyp\_h]*
- Organic crop area by agricultural production methods and crops (until 2011) [org\_cropar\_h1]*
- Share of organic crop area out of the total utilised agricultural area (UAA) (until 2011) [org\_cropar\_h2]*
- Organic crop production by crops (until 2011) [org\_croppro\_h]*
- Organic livestock (until 2011) [org\_lstspec\_h]*
- Organic production of animal products (until 2011) [org\_animal\_h]*

At the bottom of the interface is a large blue button labeled "Choose".

Σχήμα 5.11: Εισαγωγή του όρου «additives»



The interface features a top navigation bar with two buttons: "Navigate by Themes" and "Search Directly". Below this is a search input field containing the text "food additives". Below the input field, a list of search results is displayed, each preceded by a small icon and followed by a code in brackets. The results are:

- Organic operators by status of the registration process (until 2011) [org\_coptyp\_h]*
- Organic crop area by agricultural production methods and crops (until 2011) [org\_cropar\_h1]*
- Share of organic crop area out of the total utilised agricultural area (UAA) (until 2011) [org\_cropar\_h2]*
- Organic crop production by crops (until 2011) [org\_croppro\_h]*
- Organic livestock (until 2011) [org\_lstspec\_h]*
- Organic production of animal products (until 2011) [org\_animal\_h]*

At the bottom of the interface is a large blue button labeled "Choose".

Σχήμα 5.12: Εισαγωγή των όρων «food additives»

### 5.1.3 Ποσοτική Αξιολόγηση Αποτελεσμάτων

Για την ανάδειξη της απόδοσης του συστήματος με την χρήση του υβριδικού αλγορίθμου BM25 έναντι της απλής αναζήτησης βάσει τίτλου, διενεργήθηκε έλεγχος με δείγμα 15 πραγματικών ερωτημάτων. Η μετρική που χρησιμοποιήθηκε είναι η ακρίβεια στα K πρώτα αποτελέσματα (Precision@K) που υπολογίζει το ποσοστό των σχετικών συνόλων δεδομένων που εμφανίστηκαν στα πρώτα K σύνολα. Ως «σχετικότητα» ορίστηκε η άμεση θεματική ανταπόκριση, ανεξάρτητα από την παρουσία αυτούσιων των όρων στον τίτλο. Το ποσοστό προκύπτει μαθηματικά από το πλήθος των σχετικών datasets διαιρούμενο με το σύνολο των datasets που βρέθηκαν με μέγιστη το 10 λόγω της χρήσης του Precision@10. Η στήλη «Αριθμός Datasets» αναφέρεται στο πλήθος των συνόλων δεδομένων που επέστρεψε η αναζήτηση, με μέγιστο όριο πάλι το 10.

Πίνακας 5.1: Σύγκριση αποτελεσματικότητας απλής και υβριδικής BM25 αναζήτησης

| No | Όροι Αναζήτησης    | P@10<br>Απλή | Αριθμός<br>Datasets | P@10<br>BM25 | Αριθμός<br>Datasets |
|----|--------------------|--------------|---------------------|--------------|---------------------|
| 1  | GDP                | 100%         | 10                  | 100%         | 10                  |
| 2  | Unemployment       | 100%         | 10                  | 100%         | 10                  |
| 3  | Hospital beds      | 100%         | 8                   | 100%         | 10                  |
| 4  | Renewable energy   | 0%           | 0                   | 100%         | 10                  |
| 5  | Pregnancy          | 0%           | 0                   | 100%         | 10                  |
| 6  | Coal market        | 0%           | 0                   | 100%         | 3                   |
| 7  | Food additives     | 0%           | 0                   | 0%           | 0                   |
| 8  | Save energy        | 0%           | 0                   | 100%         | 10                  |
| 9  | Agriculture degree | 0%           | 0                   | 100%         | 5                   |
| 10 | Tourism population | 0%           | 0                   | 100%         | 10                  |
| 11 | Salary employee    | 100%         | 1                   | 100%         | 10                  |
| 12 | Death rate         | 100%         | 10                  | 100%         | 10                  |
| 13 | School education   | 100%         | 4                   | 100%         | 10                  |
| 14 | Tourist attraction | 0%           | 0                   | 100%         | 10                  |
| 15 | Loan depth         | 0%           | 0                   | 100%         | 1                   |

Η ενδελεχής εξέταση των αποτελεσμάτων ανέδειξε την αξία χρήσης των μεταδεδομένων ESMS και των ετικετών της EuroVoc. Σε σύνθετα ερωτήματα, το σύστημα εμφάνισε εξαιρετικά συναφή αποτελέσματα σε αντίθεση με την συμβατική αναζήτηση η οποία συχνά απέτυχε πλήρως να επιστρέψει αποτελέσματα. Ιδιαίτερο ενδιαφέρον παρουσιάζει η σύγκριση τριών παραδειγμάτων αναζήτησης που και στις δύο προσεγγίσεις πέτυχαν απόλυτη ακρίβεια (100%) στα επιστρεφόμενα σύνολα δεδομένων, αλλά η συμβατική προσέγγιση επέστρεψε μικρότερο αριθμό συνόλων. Συγκεκριμένα, τα παραδείγματα «Hospital beds», «Salary employee» και «School education» επέστρεψαν 8, 1 και 4 σύνολα δεδομένων αντίστοιχα στην συμβατική προσέγγιση, ενώ με τη χρήση του αλγορίθμου BM25 και στα τρία

παραδείγματα ανακτήθηκαν τουλάχιστον 10 σύνολα, με εξίσου απόλυτη ακρίβεια (100%). Η διαφορά αυτή φανερώνει τον σημαντικό ρόλο που έχει η ευρετηρίαση των μεταδεδομένων ESMS και του εμπλουτισμού της EuroVoc στην ακριβή αναζήτηση, επιτυγχάνοντας την ανάκτηση επιπλέον συναφών συνόλων δεδομένων, τα οποία παραλείπονται από την συμβατική προσέγγιση λόγω της απουσίας των συγκεκριμένων όρων από τον τίτλο του συνόλου δεδομένων. Συνεπώς, αυξάνεται με αυτόν τον τρόπο η ανάκληση (recall) του συστήματος χωρίς να θυσιάζεται η ακρίβεια (precision). Παράλληλα, αξιοσημείωτη είναι η συμπεριφορά του αλγορίθμου σε εξειδικευμένους όρους, όπως το «loan depth». Ενώ η απλή μέθοδος δεν επιστρέφει κανένα αποτέλεσμα, ο υβριδικός αλγόριθμος BM25 εντόπισε στοχευμένα ένα συναφές dataset, υποδηλώνοντας την ικανότητα μείωσης θορύβου της μεθόδου.

Επιπρόσθετα, η αυστηρή πολιτική φιλτραρίσματος των μηδενικών score ήταν μία κρίσιμη αρχιτεκτονική απόφαση. Σε σύνθετα ερωτήματα όπως το «food additives», το σύστημα ρυθμίστηκε ώστε να αποκλείει τα αποτελέσματα (υποψήφια datasets για εμφάνιση), αν έστω ένας από τους όρους αναζήτησης συγκεντρώσει μηδενικό score. Με την συγκεκριμένη επιλογή, προτιμήθηκε η συνειδητή περικοπή πολλών datasets τα οποία θα σχετίζονταν μερικά με τους όρους αναζήτησης, με σκοπό την αποφυγή πολλών ψευδώς θετικών αποτελεσμάτων, μειώνοντας τον θόρυβο και αυξάνοντας την ακρίβεια αναζήτησης.

Συνεπώς, αναδεικνύεται η ορθότητα τόσο της στρατηγικής φιλτραρίσματος όσο και της ανάθεσης διαφορετικών βαρών στα πεδία της βάσης δεδομένων, επιλογές που λειτουργούν συνδυαστικά και οδηγούν τον αλγόριθμο σε ακριβείς και σχετικές ανακτήσεις.

## 5.2 Σύγκριση με το Επίσημο Εργαλείο της Eurostat

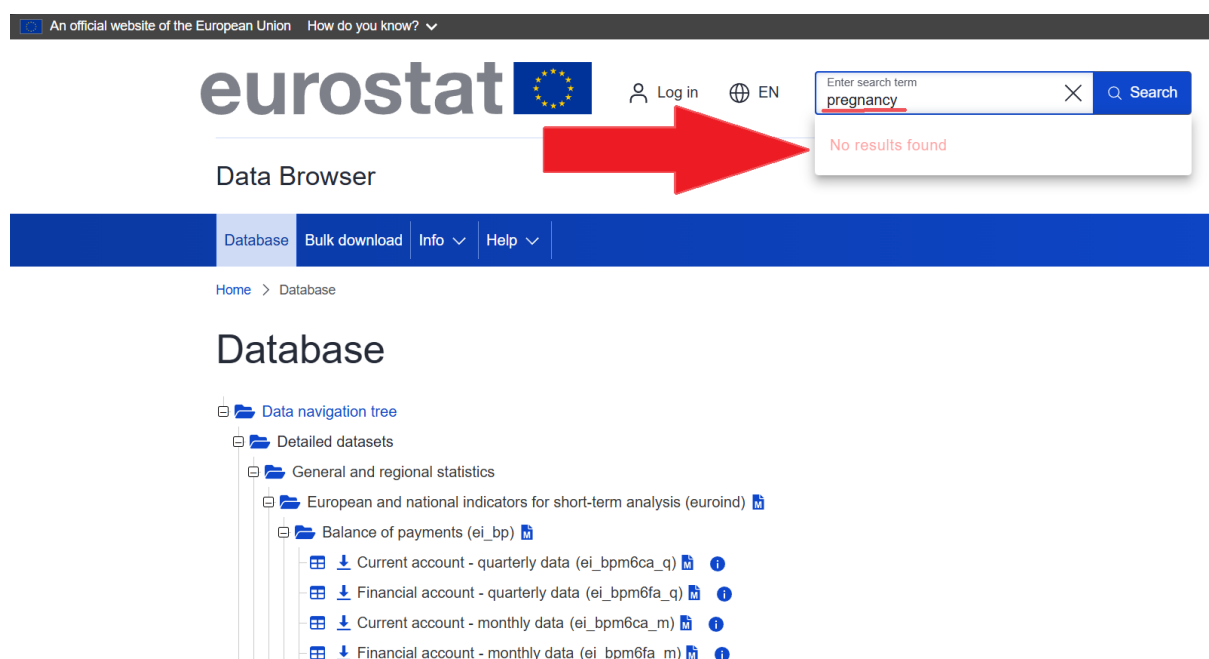
Για την πλήρη αξιολόγηση της εφαρμογής Eurostat's Finder επιχειρείται μια αντιπαραβολή της με το επίσημο εργαλείο περιήγησης δεδομένων της Eurostat. Η σύγκριση αυτή χρησιμοποιείται ενδεικτικά και αποσκοπεί κυρίως στην ανάδειξη του προβλήματος αναζήτησης και στην σημαντικότητα της ένταξης και αξιοποίησης των μεταδεδομένων σε αυτήν. Στα κοινά των δύο εφαρμογών υπάρχουν δύο τρόποι αναζήτησης dataset: α) περιήγηση στο θεματικά διαχωρισμένο ταμπλό των περιεχομένων (ToC), β) εισαγωγή όρων από τον χρήστη. Επιπλέον, εφόσον επιλεχθεί dataset τότε εμφανίζονται τα φίλτρα με τις ενεργές τιμές τους (Dimension-Codelists) και μετέπειτα εμφανίζονται τα δεδομένα σε μορφή πίνακα. Ωστόσο, στον σχεδιασμό της διεπαφής χρήστη, στην μηχανή αναζήτησης και στην αναπαράσταση των δεδομένων διαφέρουν.

### 5.2.1 Σχεδιασμός Διεπαφής Χρήστη

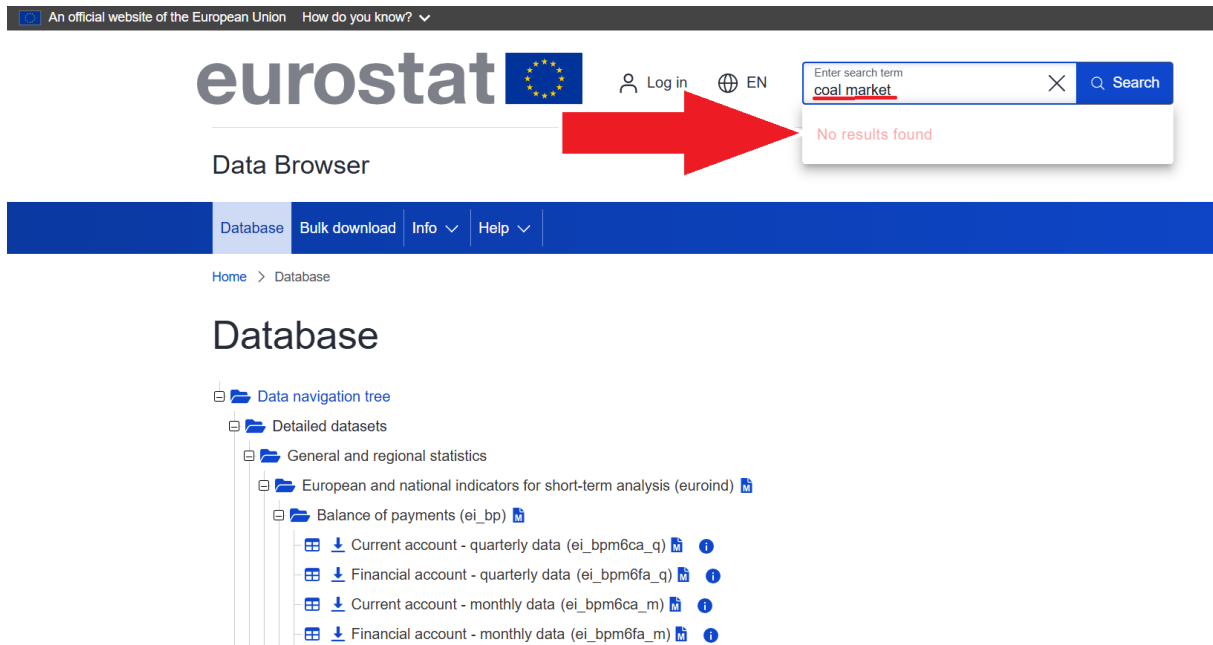
Αρχικά, το εργαλείο αναζήτησης της Eurostat διακρίνεται για την βαριά αρχιτεκτονική διεπαφής του που περιλαμβάνει πολλαπλές μεταβάσεις μεταξύ διαφορετικών σελίδων (page reloads). Συγκεκριμένα, για την εύρεση του επιθυμητού dataset φορτώνεται νέα σελίδα με τα πιθανά datasets και έπειτα από την επιλογή κάποιου dataset, φορτώνεται εκ νέου καινούρια σελίδα. Οι παραπάνω διαδικασίες δημιουργούν καθυστερήσεις και χαμηλότερου επιπέδου εμπειρίας χρήστη. Αντιθέτως, η διεπαφή χρήστη του Eurostat's Finder δημιουργήθηκε με την λογική μίας μόνο σελίδας με δυναμικές αλλαγές μέσω ασύγχρονων κλήσεων Asynchronous JavaScript and XML (AJAX) και ανάκτησης των APIs. Από το πρώτο στάδιο της εισαγωγής όρων αναζήτησης, την επιλογή dataset και μετέπειτα φίλτρων φτάνοντας τελικά στην τελική οπτικοποίηση των δεδομένων σε πίνακα, ολόκληρη η διαδικασία πραγματοποιείται αυτομάτως στην ίδια οθόνη, προσδίδοντας στην εφαρμογή μια ομαλή και σύγχρονη εμπειρία πλοήγησης.

### 5.2.2 Διαφορές στην Μηχανή Αναζήτησης

Από τις σημαντικές αδυναμίες της μηχανής αναζήτησης του επίσημου εργαλείου της Eurostat είναι η εξάρτηση ταύτισης όρων πρωτίστως με τον τίτλο και τον κωδικό των datasets κατά την πληκτρολόγηση σε πραγματικό χρόνο. Αν ένας χρήστης δεν γνωρίζει ακριβώς τον τίτλο ή το αναγνωριστικό ID, η δυναμική προβολή αδυνατεί συχνά να του προτείνει το επιθυμητό dataset, με τελευταία λύση την χειροκίνητη περιήγηση στο εκτενές δέντρο πλοήγησης (ToC). Στο Eurostat's Finder, όπως αναλύθηκε στην ενότητα 5.1, η ενσωμάτωση ετικετών EuroVoc και ο σημασιολογικός εμπλουτισμός μέσω των κειμένων μεταδεδομένων ESMS, προσφέρει σημασιολογική εμβάθυνση στην μηχανή αναζήτησης. Η υλοποίηση του αλγορίθμου BM25 πραγματοποιήθηκε έτσι ώστε να ανταποκρίνεται αποτελεσματικότερα στα σύνθετα ερωτήματα του χρήστη, καθιστώντας την αναζήτηση datasets πολύ πιο αποδοτική και διασθητική σε σχέση με το επίσημο εργαλείο. Δύο παραδείγματα εμφανίζονται στα Σχήματα 5.13 και 5.14 στα οποία δοκιμάζονται όροι εισαγωγής στο επίσημο εργαλείο της Eurostat που έχουν δοκιμαστεί παραπάνω στην εφαρμογή Eurostat's Finder και έχουν εμφανίσει συναφή αποτελέσματα. Όπως τονίζεται από τα κόκκινα βελάκια, η μηχανή αναζήτησης της Eurostat δεν εμφανίζει κανένα αποτέλεσμα σε καμία από τις δύο περιπτώσεις.



Σχήμα 5.13: Εισαγωγή όρου «pregnancy» στην μηχανή αναζήτησης επίσημου εργαλείου της Eurostat



Σχήμα 5.14: Εισαγωγή όρων «coal market» στην μηχανή αναζήτησης επίσημου εργαλείου της Eurostat

### 5.2.3 Λήψη Στατιστικών Δεδομένων

Τέλος, και οι δύο εφαρμογές δίνουν την δυνατότητα λήψης των δεδομένων. Στο επίσημο εργαλείο της Eurostat, δίνεται η δυνατότητα επιλογής λήψης με διαθέσιμες μορφές όπως JSON, XML (είτε SDMX2.1 ή SDMX3.0), TSV, CSV, XLSX. Οι μορφές αυτές (CSV, JSON, XML), παρότι είναι αναγνώσιμες από μηχανές, στερούνται ρητής σημασιολογικής αναπαράστασης και διασύνδεσης (Linked Data). Στο Eurostat's Finder όμως τα τελικά δεδομένα αναπαραστώνται στην επίσημη W3C μορφή του RDF Data Cube με σύνταξη turtle. Με αυτόν τον τρόπο γεφυρώνεται το εννοιολογικό χάσμα μεταξύ των δεδομένων και του Σημασιολογικού Ιστού αφού πλέον τα δεδομένα έχουν ήδη λάβει τις κατάλληλες ετικέτες (DCAT, SKOS, EuroVoc) και είναι έτοιμα προς άμεση και αυτόματη διαλειτουργική χρήση από μηχανές.

## 5.3 Αξιολόγηση Απόδοσης

Η αξιολόγηση του συστήματος δεν επικεντρώνεται αποκλειστικά στην σχετικότητα των αναζητήσεων και αποτελεσμάτων και στην αρχιτεκτονική της διεπαφής χρήστη, αλλά επεκτείνεται και στην τεχνική του απόδοσης. Η ανάπτυξη της εφαρμογής πραγματοποιήθηκε υπό τις αυστηρές προδιαγραφές ενός ακαδημαϊκού περιβάλλοντος κοινής φιλοξενίας. Συνεπώς, οι αρχιτεκτονικές αποφάσεις πάρθηκαν με γνώμονα τους περιορισμούς που αφορούν στον περιορισμένο αποθηκευτικό χώρο και την διαθέσιμη μνήμη επεξεργασίας.

### 5.3.1 Βελτιστοποίηση Αποθηκευτικού Χώρου

Η μεγαλύτερη τεχνική πρόκληση κατά την υλοποίηση της εφαρμογής, ήταν η αποδοτική διαχείριση του αποθηκευτικού χώρου των 250MB που παρείχε ο διακομιστής. Η αρχική προσέγγιση αποθήκευσης ολόκληρων των εκτενών μεταδεδομένων ESMS, χωρίς φιλτράρισμα, και η δοκιμή της προεπιλεγμένης αποθήκευσης InnoDB της MySQL για την βάση δεδομένων, αποδείχθηκαν μη υλοποιήσιμες. Ολόκληρα τα ακατέργαστα ESMS για αρκετά datasets ήταν πολλές σελίδες αυξάνοντας το μέγεθος του

αρχείου και μετέπειτα της βάσης δεδομένων εκθετικά, ενώ η μηχανή InnoDB δέσμευε τεράστιο χώρο αποθήκευσης για τη δημιουργία αρχείων καταγραφής (undo logs), οδηγώντας σε εξάντληση του διαθέσιμου χώρου αποθήκευσης.

Για την επίλυση του προβλήματος αποθηκευτικού χώρου, εφαρμόστηκαν οι εξής αρχιτεκτονικές επιλογές:

- **Κανονικοποίηση της βάσης δεδομένων:** Αντί να αποθηκευτεί ολόκληρο το ESMS κάθε dataset, πέρα από το φιλτράρισμα και την εξαγωγή μόνο των ενοτήτων 3.1 Description και 3.4 Concepts, τα δεδομένα χωρίστηκαν σε δύο διακριτούς πίνακες (datasets και esms\_metadata), αντί κάθε dataset να φέρει μαζί του και το ESMS κείμενο. Αξιοποιώντας την ύπαρξη του ίδιου ESMS αρχείου σε πολλαπλά datasets, εξαλείφθηκαν πλήρως τα διπλότυπα δεδομένα συρρικνώνοντας αισθητά το μέγεθος του αρχείου.
- **Μετάβαση στην μηχανή MyISAM:** Αντικατέστησε την InnoDB διότι δεν απαιτεί ογκώδη αρχεία συναλλαγών, που συνεπάγεται με αποθηκευτικό χώρο, ενώ προσφέρει βέλτιστη συμπίεση κειμένου. Με αυτόν τον τρόπο, τα κείμενα των μεταδεδομένων δεν υπέστησαν κάποια εκ νέου περικοπή, διατηρώντας ολόκληρη την σημασιολογική πληροφορία και καταλήγοντας σε βάση δεδομένων λίγο μικρότερη από 20MB.

### 5.3.2 Ταχύτητα Ανάκτησης και Υβριδική Αναζήτηση

Κατά την αναζήτηση, η εκτέλεση σύνθετων ερωτημάτων με χρήση απλών εντολών θα απαιτούσε την χρονοβόρα διαδικασία πλήρους σάρωσης των πινάκων, οδηγώντας τον εξυπηρετητή σε μεγάλες καθυστερήσεις. Για την αποφυγή αυτών, ο αλγόριθμος αναζήτησης, με χρήση της εντολής MATCH() AGAINST(), λειτουργεί πάνω σε ανεστραμμένα ευρετήρια (Inverted Indexes), αξιοποιώντας πλήρως τα ευρετήρια πλήρους κειμένου (Full Text Indexes) στις στήλες των τίτλων, θεμάτων και κειμένων ESMS. Επιπλέον, η υλοποίηση υβριδικής αναζήτησης Cross-Field AND Matching ελέγχει αν υπάρχουν πολλαπλές λέξεις-κλειδιά ταυτόχρονα στο ίδιο dataset, ακόμη και σε διαφορετικά πεδία, και εκτελείται ταχύτατα δίνοντας την δυνατότητα στον διακομιστή να υπολογίζει γρήγορα τα βάρη και να κατατάσσει σωστά τα datasets. Η εμφάνιση πιθανών dataset πραγματοποιείται με την εισαγωγή του δεύτερου γράμματος στην μηχανή αναζήτησης, αλλάζοντας δυναμικά και γρήγορα τα υποψήφια datasets ανάλογα με την αλφαριθμητική ακολουθία που εισάγει ο χρήστης. Η εξάλειψη της καθυστέρησης στο back-end είναι υπεύθυνη για την λειτουργία της διεπαφής σε πραγματικό χρόνο.

### 5.3.3 Δυναμική Παραγωγή Σημασιολογικών Δεδομένων

Σημαντική συμβολή στην απόδοση της εφαρμογής, αποτελεί η χρήση της «on-the-fly» αρχιτεκτονικής για τα RDF δεδομένα. Η προσπάθεια ανάκτησης και αποθήκευσης εκ των προτέρων όλων των στατιστικών δεδομένων της Eurostat, προϋποθέτει εκατοντάδες Gigabytes αποθηκευτικού χώρου, καθιστώντας το απίθανο για τα αυστηρά ακαδημαϊκά όρια στον δίσκο. Αντ' αυτού, ακολουθείται η αρχιτεκτονική «on-the-fly», το σύστημα κρατάει μόνον τον σημασιολογικό εμπλουτισμό στην βάση δεδομένων και αντλεί όλα τα πραγματικά δεδομένα μέσω των APIs της Eurostat. Η διαδικασία μετατροπής σε RDF Data Cube πραγματοποιείται στην μνήμη Random Access Memory (RAM) του διακομιστή κατά την διάρκεια της κλήσης, στο αρχείο convert.php. Επομένως, η εφαρμογή μπορεί να χαρακτηριστεί εξαιρετικά ελαφριά και βιώσιμη εφόσον ελαχιστοποιεί την ανάγκη για μόνιμο αποθηκευτικό χώρο (reduced storage footprint), ξεπερνώντας τους ακαδημαϊκούς περιορισμούς.

## 5.4 Ποσοτική Αποτίμηση και Μετρικές Συστήματος

Για την ολική αποτίμηση των βελτιώσεων αποθηκευτικού χώρου και υπολογιστικών πόρων, κρίνεται αναγκαία η καταμέτρηση και παράθεση ποσοτικών μετρικών που συνδέονται με τον όγκο των ανακτημένων δεδομένων, από την πρώτη ανάκτηση μέχρι το πέρας της επεξεργασίας τους.

Τα παρακάτω στατιστικά στοιχεία προέκυψαν από την εκτέλεση του αρχείου index-builder.js, υπεύθυνο για την ανάκτηση του πλήρη καταλόγου της Eurostat:

- **Συνολικό Πλήθος Datasets:** Καταγράφηκαν περίπου 9.000 ενεργά datasets με δεδομένα.
- **Συνολικό Πλήθος Μοναδικών Εγχειριδίων Μεταδεδομένων ESMS:** Ανακτήθηκαν μόνο 327 μοναδικά αρχεία ESMS, τονίζοντας την τεράστια απόκλιση που έχουν με το σύνολο των datasets και επιβεβαιώνοντας την αρχιτεκτονική κανονικοποίησης της βάσης δεδομένων που επιλέχθηκε.
- **Σύνολο Όγκου Ακατέργαστων Δεδομένων:** Τα δεδομένα των DSDs, τα Codelists και ολόκληρα τα εγχειρίδια ESMS αν δεν αποθηκεύονταν τοπικά, ο όγκος τους θα ξεπερνούσε κατά πολύ τα όρια του διακομιστή και θα εισήγαγε παραπάνω στατιστικό θόρυβο.
- **Σύνολο Μεγέθους Βάσης Δεδομένων :** Η επιλογή της μηχανής MyISAM, ο διαχωρισμός των πινάκων και το έξυπνο φιλτράρισμα θορύβου είχε ως αποτέλεσμα την συμπίεση των δεδομένων σε μέγεθος μικρότερο των 20 MB. Βελτιώνοντας αισθητά τις καθυστερήσεις.

Η μείωση του αποθηκευτικού αποτυπώματος κατά σχεδόν 90% αποτελεί δείκτη απόλυτης επιτυχίας, καθώς επιτρέπει την ταχύτατη φόρτωση του ευρετηρίου στη μνήμη RAM κατά τις αναζητήσεις, ελαχιστοποιώντας τον χρόνο ανάκτησης, ακόμα και σε ερωτήματα πολλαπλών λέξεων.

Προκειμένου να αποτιμηθεί ποσοτικά η ταχύτητα της εφαρμογής σε πραγματικό χρόνο, μετρήθηκαν οι χρόνοι απόκρισης του συστήματος σε δύο τυχαία παραδείγματα («car» και «traffic lights») στις τρεις βασικές του φάσεις: κατά την αναζήτηση των συνόλων δεδομένων, κατά την φόρτωση των φίλτρων και, τέλος, κατά την ανάκτηση των δεδομένων.

Πίνακας 5.2: Χρόνοι απόκρισης βασικών λειτουργιών της Eurostat's Finder

| Λειτουργία Συστήματος | Χρόνος Εκτέλεσης για «car» | Χρόνος Εκτέλεσης για «traffic lights» |
|-----------------------|----------------------------|---------------------------------------|
| Αναζήτηση             | 71ms                       | 70ms                                  |
| Ανάκτηση Φίλτρων      | 105ms                      | 6,05sec                               |
| Ανάκτηση Δεδομένων    | 554ms                      | 533ms                                 |

Όπως παρατηρείται, ο σχεδιασμός του ενδιάμεσου τοπικού ευρετηρίου και η απουσία περιττών φορτώσεων (page reloads), διασφαλίζουν την άμεση ανταπόκριση του συστήματος σε κλάσματα του δευτερολέπτου. Μοναδική εξαίρεση αποτελεί το στάδιο ανάκτησης φίλτρων, όπου οι χρόνοι ενδέχεται να φτάσουν τα μερικά δευτερόλεπτα. Αυτό συμβαίνει λόγω της εξάρτησης της ανάκτησης των φίλτρων με τον χρόνο απόκρισης των εξωτερικών APIs της Eurostat και είναι ανάλογη του πλήθους των διαστάσεων και των αποδεκτών τιμών (Codelists) που υποστηρίζει το εκάστοτε επιλεγμένο dataset. Συνεπώς είναι λογικό να παρατηρούνται διακυμάνσεις στον χρόνο απόκρισης ανάλογα με τον όγκο πληροφορίας που καλείται να επιστρέψει ο εξωτερικός διακομιστής.

## Κεφάλαιο 6ο: Συμπεράσματα και Προτάσεις Βελτίωσης

### 6.1 Σύνοψη και Συμπεράσματα

Η παρούσα διπλωματική εργασία δημιουργήθηκε από τις αδυναμίες στο ήδη υπάρχον επίσημο εργαλείο της Eurostat και επικεντρώθηκε στον σχεδιασμό, την υλοποίηση και την αξιολόγηση της εφαρμογής Eurostat's Finder, που κατασκευάστηκε για να επιλύσει τις υπάρχουσες αδυναμίες. Βασικότερος στόχος ήταν η μετατροπή της αναζήτησης, από χειροκίνητη σε σημασιολογικά εμπλουτισμένη, διευκολύνοντας τους τελικούς χρήστες και η σημασιολογική αναπαράσταση και αξιοποίηση των εξαγομένων στατιστικών δεδομένων.

Έπειτα από συνεχή έρευνα και πρακτικές δοκιμές και υλοποιήσεις, εξήχθησαν σημαντικά συμπεράσματα σχετικά με την διαχείριση των δεδομένων και με την αρχιτεκτονική του λογισμικού. Αρχικά, επιβεβαιώθηκε ότι ο σημασιολογικός εμπλουτισμός μέσω των EuroVoc ετικετών και την προσθήκη των ESMS μεταδεδομένων, βελτίωσε την ακρίβεια της αναζήτησης εμφανίζοντας θεματικά συναφή datasets τα οποία δεν ανακτώνταν με τη χρήση συμβατικών μεθόδων. Η υλοποίηση του αλγορίθμου κατανομής βαρών, υβριδική μορφή του αλγορίθμου BM25, επέτρεψε την αποδοτική κατανομή και εμφάνιση των datasets με σειρά βαθμολόγησης συνάφειας, με τα datasets να εμφανίζονται βάσει φθίνουσας βαθμολογίας (score). Ενώ η προσέγγιση Cross-Field AND Matching περιόρισε την ύπαρξη ψευδών θετικών (false positives) στις περιπτώσεις εισαγωγής πολλαπλών όρων αναζήτησης.

Σε αρχιτεκτονικό επίπεδο, η εργασία ανέδειξε τη σημασία της προσαρμοστικότητας βάσει περιορισμών σε περιβάλλοντα φιλοξενίας. Η επιλογή αποθήκευσης JSON αρχείων σε σχεσιακή βάση δεδομένων MySQL με χρήση της μηχανής MyISAM, ξεπέρασε τους τεχνικούς περιορισμούς φιλοξενίας στον αυστηρά περιορισμένο αποθηκευτικό χώρο. Επίσης, το έξυπνο φιλτράρισμα ενοτήτων των αρχείων ESMS, έδωσε την δυνατότητα εισαγωγής δύο ολόκληρων ενοτήτων, της περιγραφής και των εννοιών. Η εισαγωγή τους, χωρίς περικοπές, κρίνεται υψίστης σημασίας καθώς προσδίδουν μεγάλο μέρος του σημασιολογικού εμπλουτισμού ενώ με την μηχανή MyISAM επιτρέπεται η ταχύτατη αναζήτηση μες τα εκτενή κείμενά τους.

Τέλος, το πιο καινοτόμο συμπέρασμα αφορά την αρχιτεκτονική «on-the-fly» και την χρησιμότητα της τεχνολογίας RDF στην αναπαράσταση των στατιστικών δεδομένων. Πραγματοποιήθηκε πλήρης δυναμική μετατροπή σε πραγματικό χρόνο, των ανακτώμενων δεδομένων μέσω των Eurostat APIs στην επίσημη μορφή του W3C RDF Data Cube, παράγοντας παράλληλα λεξιλόγιο SKOS για την ερμηνεία των κωδικών. Με αυτόν τον τρόπο αποφεύχθηκε τεράστιος αριθμός τοπικών bytes για την διάθεση των δεδομένων ως Linked Open Data. Συνεπώς, η εφαρμογή Eurostat's Finder, ελαχιστοποίησε το αποτύπωμα μόνιμης αποθήκευσης των δεδομένων ενώ ταυτόχρονα διατήρησε την μηχανική της διαλειτουργικότητα.

#### 6.1.1 Περιορισμοί Υφιστάμενης Υλοποίησης

Παρά τα θετικά αποτελέσματα της αξιολόγησης, υπάρχουν περιορισμοί στην τρέχουσα αρχιτεκτονική που χρήζουν αναφοράς. Καταρχάς, η ποσοτική αξιολόγηση (Precision) εφαρμόστηκε πάνω στο πεπερασμένο πλήθος των 15 ερωτημάτων. Παρότι τα παραδείγματα που επιλέχθηκαν ήταν αντιπροσωπευτικά του συνόλου, το δείγμα παραμένει μικρό για ασφαλή εξαγωγή καθολικών συμπερασμάτων. Επιπλέον, ο σημασιολογικός εμπλουτισμός που εφαρμόζει το σύστημα πρόκειται για αντιστοίχιση ετικετών με το λεξιλόγιο του EuroVoc και όχι πλήρη σημασιολογική συλλογιστική. Όσον αφορά την εξαγωγή του μετρήσιμου μεγέθους `observedMeasureLabel`, αυτή βασίζεται σε ευρετικούς

κανόνες λεκτικής αποκοπής και λειτουργεί ικανοποιητικά σε μεγάλο πλήθος datasets αλλά δεν αποτελεί καθολική λύση. Τέλος, η τρέχουσα έκδοση της εφαρμογής λειτουργεί αποκλειστικά στην αγγλική γλώσσα και δεν παρέχει ανοιχτό SPARQL endpoint για προγραμματιστική (machine-to-machine) επερώτηση των RDF δεδομένων από τρίτες εφαρμογές.

## 6.2 Μελλοντικές Επεκτάσεις

Παρά την επίτευξη των στόχων που τέθηκαν η εφαρμογή Eurostat's Finder, αρκετές βελτιώσεις μπορούν να εφαρμοστούν σε ήδη υπάρχοντα εργαλεία αλλά και νέες ιδέες επέκτασης της εφαρμογής και σε άλλους τομείς. Παρακάτω παρουσιάζονται οι εξής επεκτάσεις:

### 6.2.1 Ενσωμάτωση Μηχανισμών Επεξεργασίας Φυσικής Γλώσσας

Στην τρέχουσα υλοποίηση της εφαρμογής, τα ευρετήρια πλήρους κειμένου που ανακτήθηκαν από τα εγχειρίδια ESMS, λαμβάνονται ως έχουν χωρίς καμία μορφοποίηση. Παρότι στο θεωρητικό υπόβαθρο αναλύθηκαν οι τεχνικές Επεξεργασίας Φυσικής Γλώσσας (NLP) και μερικές χρησιμοποιήθηκαν στην εφαρμογή, η τεχνική της λημματοποίησης (lemmatization) δεν εφαρμόστηκε στην τωρινή έκδοση της. Συνεπώς μια μελλοντική επέκταση θα μπορούσε να ενσωματώνει σωστά τους αλγόριθμους της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing - NLP) κατά το στάδιο της προετοιμασίας των δεδομένων, θα οδηγούσε σε πιο ακριβή αποτελέσματα του υβριδικού αλγορίθμου BM25.

### 6.2.2 Δυναμικός Αλγόριθμος Εντοπισμού Μετρήσιμου Μεγέθους

Όπως διαπιστώθηκε κατά την ανάλυση των DSDs, τα σύνολα δεδομένων της Eurostat εμφανίζουν έλλειψη μιας γενικής, τυποποιημένης ετικέτας που να ορίζει ρητά το μετρήσιμο μέγεθος τους κάθε dataset. Σε ορισμένα datasets, η πληροφορία του μετρήσιμου μεγέθους βρίσκεται ενσωματωμένη μες τον τίτλο, ενώ σε άλλα υπάρχει συγκεκριμένη διάσταση (όπως η indic), που ανάλογα με την τιμή που θέτει ο χρήστης, αλλάζει αντίστοιχα και το μετρήσιμο μέγεθος.

Στην παρούσα έκδοση, δόθηκε μία προσαρμοσμένη λύση στο πρόβλημα της εξαγωγής μετρήσιμου μεγέθους – το οποίο είναι απαραίτητο συστατικό για την ορθή δόμηση του RDF Data Cube. Η λύση αυτή βασίστηκε σε ευρετικούς μηχανισμούς που καλύπτουν μερικά datasets αλλά δεν αποτελεί ολοκληρωμένη λύση αφού περιορίζει σε έναν βαθμό την άμεση διαλειτουργικότητα του RDF Data Cube Vocabulary, από την εισαγωγή προσαρμοσμένων ετικετών (πρόθεμα estatfinder:). Επομένως, μελλοντικά θα μπορούσε να υλοποιηθεί νέος αλγόριθμος εύρεσης μετρήσιμου μεγέθους που θα αποδίδει αποδοτικά και καθολικά την έννοια αυτήν. Με αυτόν τον τρόπο θα καταργήσει την ανάγκη χρήσης προσαρμοσμένων ετικετών, συμβαδίζοντας πλήρως με τα αυστηρά πρότυπα του W3C RDF Data Cube.

### 6.2.3 Παροχή Δημόσιου REST API

Στην παρούσα έκδοση, η εφαρμογή καθιστά δυνατή την παράδοση RDF αρχείου αποκλειστικά μέσω λήψης, χρησιμοποιώντας το UI της εφαρμογής. Μελλοντικά, θα μπορούσε να υπάρξει επέκταση της παροχής αρχείων RDF μέσω αυτόνομου RESTful APIs. Η δημιουργία αυτή, θα καταργήσει την ανάγκη ύπαρξης ανθρώπινης παρέμβασης για την αποστολή ερωτημάτων από εξωτερικά λογισμικά και πράκτορες λογισμικού, αφού πλέον τα ερωτήματα θα μπορούν να στέλνονται αυτοματοποιημένα στο αυτόνομο API.

#### 6.2.4 Υποστήριξη Άλλων Γλωσσών

Το επίσημο εργαλείο της Eurostat, μέσω των APIs του, δίνει την δυνατότητα διανομής της πληροφορίας σε τρεις γλώσσες; Αγγλικά, Γερμανικά, Γαλλικά. Στην εφαρμογή χρησιμοποιήθηκε μόνο η αγγλική γλώσσα ως η πιο κοινή και διαδεδομένη. Όμως μια μελλοντική επέκταση της εφαρμογής, θα μπορούσε να δίνει την δυνατότητα στον χρήστη να επιλέξει σε ποια γλώσσα θέλει να περιηγηθεί. Επίσης, υπάρχει και η δυνατότητα μετάφρασης γλωσσών στις 24 γλώσσες της ΕΕ, που τις υποστηρίζει και ο θησαυρός της EuroVoc, αλλάζοντας όλο το κείμενο του UI και όλα τα δεδομένα, φίλτρα και κείμενα της Eurostat προσφέροντας πραγματικά μια διασυνοριακή εμπειρία χρήσης.

#### 6.2.5 Αυτοματοποιημένος Μηχανισμός Συγχρονισμού Δεδομένων

Η Eurostat ανανεώνει τα δεδομένα της δύο φορές την ημέρα, μία στις 11:00 και μία στις 23:00. Η ανανέωση αυτή αφορά είτε τροποποίηση ήδη υπάρχοντος dataset είτε προσθήκη νέου dataset εξ ολοκλήρου. Συνεπώς, δεδομένα εισέρχονται/μεταβάλλονται καθημερινά στην Eurostat, τονίζοντας την ανάγκη ύπαρξης αυτοματοποιημένου μηχανισμού για την συλλογή των στατιστικών πληροφοριών. Στην παρούσα κατάσταση της εφαρμογής, η διαδικασία συλλογής των πληροφοριών και έπειτα δημιουργίας του τοπικού ευρετηρίου πραγματοποιείται χειροκίνητα και εκτός σύνδεσης.

Μια σημαντική μελλοντική επέκταση θα ήταν η αυτοματοποίηση του μηχανισμού συγχρονισμού συλλογής δεδομένων. Χρησιμοποιώντας προγραμματισμένες διαδικασίες (όπως Cron jobs) βάσει του ωραρίου ενημέρωσης της Eurostat, το σύστημα θα μπορούσε να εκτελεί αυτόματα τα scripts ανάκτησης δεδομένων. Με αυτόν τον τρόπο, οι αλλαγές στον κατάλογο ToC, στα μεταδεδομένα ESMS και στους σημασιολογικούς όρους εμπλουτισμού της EuroVoc, θα ενημερώνονται αυτόματα όταν συμβαίνουν οι ενημερώσεις και μετέπειτα θα ενημερώνονται και οι πίνακες της βάσης δεδομένων, χωρίς να προκαλείται διακοπή στις υπηρεσίες προς τους χρήστες. Με αυτήν την αλλαγή, το Eurostat's Finder αναβαθμίζεται σε μία εφαρμογή πλήρως αυτόνομη που διασφαλίζει την εγκυρότητα και την επικαιρότητα των στοιχείων της.

# ΒΙΒΛΙΟΓΡΑΦΙΑ

## Internet Site

- [1] Open Knowledge Foundation, "The Open Data Handbook," [Online]. Available: <https://opendatahandbook.org/>.
- [2] European Parliament and Council of the European Union, "Directive (EU) 2019/1024 on open data and the re-use of public sector information," Official Journal of the European Union, 2019. [Online]. Available: <https://eur-lex.europa.eu/eli/dir/2019/1024/oj>.
- [4] Eurostat, "SDMX explained," European Commission. [Online]. Available: <https://ec.europa.eu/eurostat/web/sdmx-infospace/sdmx-explained>.
- [5] SDMX, "What is SDMX?," SDMX.org. [Online]. Available: <https://sdmx.org/about-sdmx/welcome/>.
- [6] Eurostat, "Metadata - Euro SDMX Metadata Structure (ESMS)," European Commission. [Online]. Available: <https://ec.europa.eu/eurostat/data/metadata/metadata-structure>.
- [8] W3C, "Resource Description Framework (RDF)," W3C Recommendation. [Online]. Available: <https://www.w3.org/RDF/>.
- [9] W3C, "SPARQL 1.1 Query Language," W3C Recommendation, Mar. 2013. [Online]. Available: <https://www.w3.org/TR/sparql11-query/>.
- [10] W3C, "SKOS Simple e Organization System Reference," W3C Recommendation, 2009. [Online]. Available: <https://www.w3.org/TR/skos-reference/>.
- [11] Publications Office of the European Union, "Thesauri," EU Vocabularies. [Online]. Available: <https://op.europa.eu/en/web/eu-vocabularies/thesauri>.
- [12] W3C, "The RDF Data Cube Vocabulary," W3C Recommendation, Jan. 2014. [Online]. Available: <https://www.w3.org/TR/vocab-data-cube/>.
- [16] Attard, Judie & Orlandi, Fabrizio & Scerri, Simon & Auer, Sören. (2015). A Systematic Review of Open Government Data Initiatives. *Government Information Quarterly*. 32. 10.1016/j.giq.2015.07.006.
- [17] Hannah Bast, Björn Buchhold, and Elmar Haussmann. 2016. Semantic Search on Text and Knowledge Bases. *Found. Trends Inf. Retr.* 10, 2–3 (Jun 2016), 119–271.
- [18] P. E. R. Salas, M. Martin, F. M. D. Mota, S. Auer, K. Breitman and M. A. Casanova, "Publishing Statistical Data on the Web," 2012 IEEE Sixth International Conference on Semantic Computing, Palermo, Italy, 2012, pp. 285-292.
- [21] Tykhonov, V. (2023, November 3). Semantic search in data repositories. Zenodo. ODISSEI Conference for Social Science in the Netherlands (OCSSN), Utrecht, the Netherlands.
- [25] Kakitaeva, Meerim & Gaso, Mekia. (2025). Cross-Origin Resource Sharing (CORS) Policy Enforcement in Spring Boot: Security Implications and Best Practices. 10.20944/preprints202505.1312.v1.
- [27] W3C, "Data Catalog Vocabulary (DCAT) - Version 2", W3C Recommendation, Feb. 2020. [Online]. Available: <https://www.w3.org/TR/vocab-dcat-2/>.

[28] DCMI Usage Board, "DCMI Metadata Terms", Dublin Core Metadata Initiative, Jan. 2020. [Online]. Available: <https://dublincore.org/specifications/dublin-core/dcmi-terms/>.

### **Journal Articles**

[3] A. Zuiderwijk, M. Janssen, S. Choenni, R. Meisner, and A. Sheikhan, "Socio-technical impediments of open data," *Journal of e-Government*, vol. 10, no. 2, pp. 156-172, 2012.

[7] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, vol. 284, no. 5, pp. 34-43, 2001.

[15] Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4), 258–268.

[19] Barati, M. (2023). Open Government Data Programs and Information Privacy Concerns: A Literature Review. *JeDEM - EJournal of EDemocracy and Open Government*, 15(1), 73–123.

[20] Ishengoma F, Shao D (2025), "Ethical concerns of open government data". *Journal of Ethics in Entrepreneurship and Technology*, Vol. 5 No. 2 pp. 206–222.

[22] Rajabi, Enayat & Midha, Rishi & Souza, Jairo. (2023). Constructing a knowledge graph for open government data: the case of Nova Scotia disease datasets. *Journal of Biomedical Semantics*. 14. 10.1186/s13326-023-00284-w.

[23] Papadaki, M.-E.; Tzitzikas, Y.; Mountantonakis, M. A Brief Survey of Methods for Analytics over RDF Knowledge Graphs. *Analytics* 2023, 2, 55-74.

[24] Tran et al., (2026). FAIRLinked: Data FAIRification Tools for Materials Data Science. *Journal of Open Source Software*, 11(122), 9628.

[26] J. Chen, J. Jiang, H. Duan, T. Wan, S. Chen, V. Paxson, and M. Yang, "We Still Don't Have Secure Cross-Domain Requests: an Empirical Study of CORS," in *27th USENIX Security Symposium (USENIX Security 18)*, pp. 1079-1093, 2018.

### **Βιβλία**

[13] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[14] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

## ΠΑΡΑΡΤΗΜΑ Α : Πλήρες Παράδειγμα Παραγόμενου Αρχείου Σημασιολογικού Κύβου RDF Data Cube

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .  
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .  
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .  
@prefix dcterms: <http://purl.org/dc/terms/> .  
@prefix dcat: <http://www.w3.org/ns/dcat#> .  
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .  
@prefix qb: <http://purl.org/linked-data/cube#> .  
@prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#> .  
@prefix sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#> .  
@prefix sdmx-attribute: <http://purl.org/linked-data/sdmx/2009/attribute#> .  
@prefix estat: <http://eurostat.linked-statistics.org/> .  
@prefix estatdim: <http://eurostat.linked-statistics.org/dimension/> .  
@prefix estatcode: <http://eurostat.linked-statistics.org/code/> .

# Dataset

estat:dataset\_nama\_10\_gdp a qb:DataSet ;

dcterms:identifier "nama\_10\_gdp" ;

dcterms:title "Gross domestic product (GDP) and main components (output, expenditure and income)  
- annual data"@en ;

rdfs:label "Gross domestic product (GDP) and main components (output, expenditure and income) -  
annual data"@en ;

dcterms:description "National accounts are a coherent and consistent set of macroeconomic indicators, which provide an overall picture of the economic situation and are widely used for economic analysis and forecasting, policy design and policy making. Annual national accounts are compiled in accordance with the European System of Accounts - ESA 2010 as defined in Annex B of the Council Regulation (EU) No 549/2013 of the European Parliament and of the Council of 21 May 2013, amended by Council Regulation (EU) 2023/734 of 15 March 2023. Gross domestic product (GDP) is one of the key aggregates in the European system of accounts (ESA). GDP is a measure of the total economic activity taking place on an economic territory which leads to output meeting the final demands of the economy. There are three ways of measuring GDP at market prices: the production approach, as the sum of the values added by all activities which produce goods and services, plus taxes less subsidies on products; the expenditure approach, as the total of all final expenditures made in either consuming the final output of the economy, or in adding to wealth, plus exports less imports of goods and services; the

income approach, as the total of all incomes earned in the process of producing goods and services plus taxes on production and imports less subsidies. Data published in the following tables reflect these 3 approaches. Breakdowns provided are based on the ESA Transmission Programme, which list all tables requested from the countries. The annual tables under this collection are the following: nama\_10\_gdp GDP and main components (output, expenditure and income) nama\_10\_pc Main GDP aggregates per capita nama\_10\_a10 Gross value added and income components by A\*10 industry nama\_10\_a64 Gross value added and income components by A\*64 industry Geographical entities covered are the European Union, the euro area, EU Member States, EFTA countries and Candidate Countries. Data from other countries (e.g. US, Japan and other countries) are received via the OECD and IMF and published in Eurobase in the naid\_10 collection. Data sources: National Statistical Institutes, Eurostat (for European aggregates)."@en ;

dc:theme "General and regional statistics"@en ;  
 dc:keyword "non-profit institutions serving households"@en ;  
 dc:keyword "government final consumption expenditure"@en ;  
 dc:keyword "european system of accounts 2010"@en ;  
 dc:keyword "general and regional statistics"@en ;  
 dc:keyword "gross fixed capital formation"@en ;  
 dc:keyword "european system of accounts"@en ;  
 dc:keyword "intermediate consumption"@en ;  
 dc:keyword "consumption expenditure"@en ;  
 dc:keyword "gross domestic product"@en ;  
 dc:keyword "candidate countries"@en ;  
 dc:keyword "economic indicators"@en ;  
 dc:keyword "european parliament"@en ;  
 dc:keyword "regional statistics"@en ;  
 dc:keyword "council regulation"@en ;  
 dc:keyword "economic indicator"@en ;  
 qb:structure estat:dsd\_nama\_10\_gdp .

#### # Observations

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-1995 a  
 qb:Observation ;  
 qb:dataSet estat:dataset\_nama\_10\_gdp ;  
 sdmx-dimension:timePeriod "1995"^^xsd:gYear ;  
 sdmx-measure:obsValue "89.114"^^xsd:decimal ;

estatdim:FREQ estatcode:FREQ\_A ;  
 estatdim:GEO estatcode:GEO\_EL ;  
 estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
 estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
 .

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-1996 a  
 qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
 sdmx-dimension:timePeriod "1996"^^xsd:gYear ;  
 sdmx-measure:obsValue "89.859"^^xsd:decimal ;  
 estatdim:FREQ estatcode:FREQ\_A ;  
 estatdim:GEO estatcode:GEO\_EL ;  
 estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
 estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
 .

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-1997 a  
 qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
 sdmx-dimension:timePeriod "1997"^^xsd:gYear ;  
 sdmx-measure:obsValue "92.828"^^xsd:decimal ;  
 estatdim:FREQ estatcode:FREQ\_A ;  
 estatdim:GEO estatcode:GEO\_EL ;  
 estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
 estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
 .

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-1998 a  
 qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
 sdmx-dimension:timePeriod "1998"^^xsd:gYear ;  
 sdmx-measure:obsValue "96.760"^^xsd:decimal ;  
 estatdim:FREQ estatcode:FREQ\_A ;

estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-1999                      a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "1999"^^xsd:gYear ;  
sdmx-measure:obsValue "98.148"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2000                      a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2000"^^xsd:gYear ;  
sdmx-measure:obsValue "102.053"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2001                      a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2001"^^xsd:gYear ;  
sdmx-measure:obsValue "106.076"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;

estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;

estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2002      a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;

sdmx-dimension:timePeriod "2002"^^xsd:gYear ;

sdmx-measure:obsValue "111.074"^^xsd:decimal ;

estatdim:FREQ estatcode:FREQ\_A ;

estatdim:GEO estatcode:GEO\_EL ;

estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;

estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2003      a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;

sdmx-dimension:timePeriod "2003"^^xsd:gYear ;

sdmx-measure:obsValue "117.911"^^xsd:decimal ;

estatdim:FREQ estatcode:FREQ\_A ;

estatdim:GEO estatcode:GEO\_EL ;

estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;

estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2004      a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;

sdmx-dimension:timePeriod "2004"^^xsd:gYear ;

sdmx-measure:obsValue "125.218"^^xsd:decimal ;

estatdim:FREQ estatcode:FREQ\_A ;

estatdim:GEO estatcode:GEO\_EL ;

estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;

estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2005 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;

sdmx-dimension:timePeriod "2005"^^xsd:gYear ;

sdmx-measure:obsValue "126.554"^^xsd:decimal ;

estatdim:FREQ estatcode:FREQ\_A ;

estatdim:GEO estatcode:GEO\_EL ;

estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;

estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2006 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;

sdmx-dimension:timePeriod "2006"^^xsd:gYear ;

sdmx-measure:obsValue "133.191"^^xsd:decimal ;

estatdim:FREQ estatcode:FREQ\_A ;

estatdim:GEO estatcode:GEO\_EL ;

estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;

estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2007 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;

sdmx-dimension:timePeriod "2007"^^xsd:gYear ;

sdmx-measure:obsValue "137.132"^^xsd:decimal ;

estatdim:FREQ estatcode:FREQ\_A ;

estatdim:GEO estatcode:GEO\_EL ;

estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;

estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.  
estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2008 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2008"^^xsd:gYear ;  
sdmx-measure:obsValue "137.535"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.  
estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2009 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2009"^^xsd:gYear ;  
sdmx-measure:obsValue "133.015"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.  
estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2010 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2010"^^xsd:gYear ;  
sdmx-measure:obsValue "125.256"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2011 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2011"^^xsd:gYear ;  
sdmx-measure:obsValue "114.140"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2012 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2012"^^xsd:gYear ;  
sdmx-measure:obsValue "105.924"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2013 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2013"^^xsd:gYear ;  
sdmx-measure:obsValue "104.160"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;

.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2014 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2014"^^xsd:gYear ;  
sdmx-measure:obsValue "104.280"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2015 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2015"^^xsd:gYear ;  
sdmx-measure:obsValue "104.288"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2016 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2016"^^xsd:gYear ;  
sdmx-measure:obsValue "103.926"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2017 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2017"^^xsd:gYear ;  
sdmx-measure:obsValue "106.260"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2018 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2018"^^xsd:gYear ;  
sdmx-measure:obsValue "108.094"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2019 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2019"^^xsd:gYear ;  
sdmx-measure:obsValue "110.009"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2020 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2020"^^xsd:gYear ;  
sdmx-measure:obsValue "100.000"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2021 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2021"^^xsd:gYear ;  
sdmx-measure:obsValue "108.125"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2022 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2022"^^xsd:gYear ;  
sdmx-measure:obsValue "113.587"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2023 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2023"^^xsd:gYear ;  
sdmx-measure:obsValue "116.156"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2024 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2024"^^xsd:gYear ;  
sdmx-measure:obsValue "118.105"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

estat:obs\_nama\_10\_gdp\_FREQ-A\_GEO-EL\_NA\_ITEM-B1G\_UNIT-CLV\_I20\_time-2025 a  
qb:Observation ;

qb:dataSet estat:dataset\_nama\_10\_gdp ;  
sdmx-dimension:timePeriod "2025"^^xsd:gYear ;  
sdmx-measure:obsValue "119.693"^^xsd:decimal ;  
estatdim:FREQ estatcode:FREQ\_A ;  
estatdim:GEO estatcode:GEO\_EL ;  
estatdim:NA\_ITEM estatcode:NA\_ITEM\_B1G ;  
estatdim:UNIT estatcode:UNIT\_CLV\_I20 ;  
.

# Code Labels

estatcode:FREQ\_A a skos:Concept ;

skos:notation "A" ;

rdfs:label "Annual"@en ;

skos:prefLabel "Annual"@en .

estatcode:GEO\_EL a skos:Concept ;

skos:notation "EL" ;

rdfs:label "Greece"@en ;

skos:prefLabel "Greece"@en .

estatcode:NA\_ITEM\_BIG a skos:Concept ;

skos:notation "BIG" ;

rdfs:label "Value added, gross"@en ;

skos:prefLabel "Value added, gross"@en .

estatcode:UNIT\_CLV\_I20 a skos:Concept ;

skos:notation "CLV\_I20" ;

rdfs:label "Chain linked volumes, index 2020=100"@en ;

skos:prefLabel "Chain linked volumes, index 2020=100"@en .

## ΠΑΡΑΡΤΗΜΑ Β : Παραγόμενη Δομή DSD

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

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

@prefix qb: <http://purl.org/linked-data/cube#> .

@prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#> .

@prefix sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#> .

@prefix estat: <http://eurostat.linked-statistics.org/> .

@prefix estatdim: <http://eurostat.linked-statistics.org/dimension/> .

@prefix estatfinder: <http://eurostat.linked-statistics.org/finder/ontology#> .

estat:dsd\_nama\_10\_gdp a qb:DataStructureDefinition ;

    rdfs:label "Data Structure Definition for nama\_10\_gdp"@en ;

    estatfinder:observedMeasureLabel "Gross domestic product (GDP) and main components (output, expenditure and income)"@en ;

    estatfinder:measureSelectorDimension estatdim:NA\_ITEM ;

    qb:component [ qb:measure sdmx-measure:obsValue ] ;

    qb:component [ qb:dimension estatdim:FREQ ] ;

    qb:component [ qb:dimension estatdim:UNIT ] ;

    qb:component [ qb:dimension estatdim:NA\_ITEM ] ;

    qb:component [ qb:dimension estatdim:GEO ] ;

    qb:component [ qb:dimension sdmx-dimension:timePeriod ] ;

.