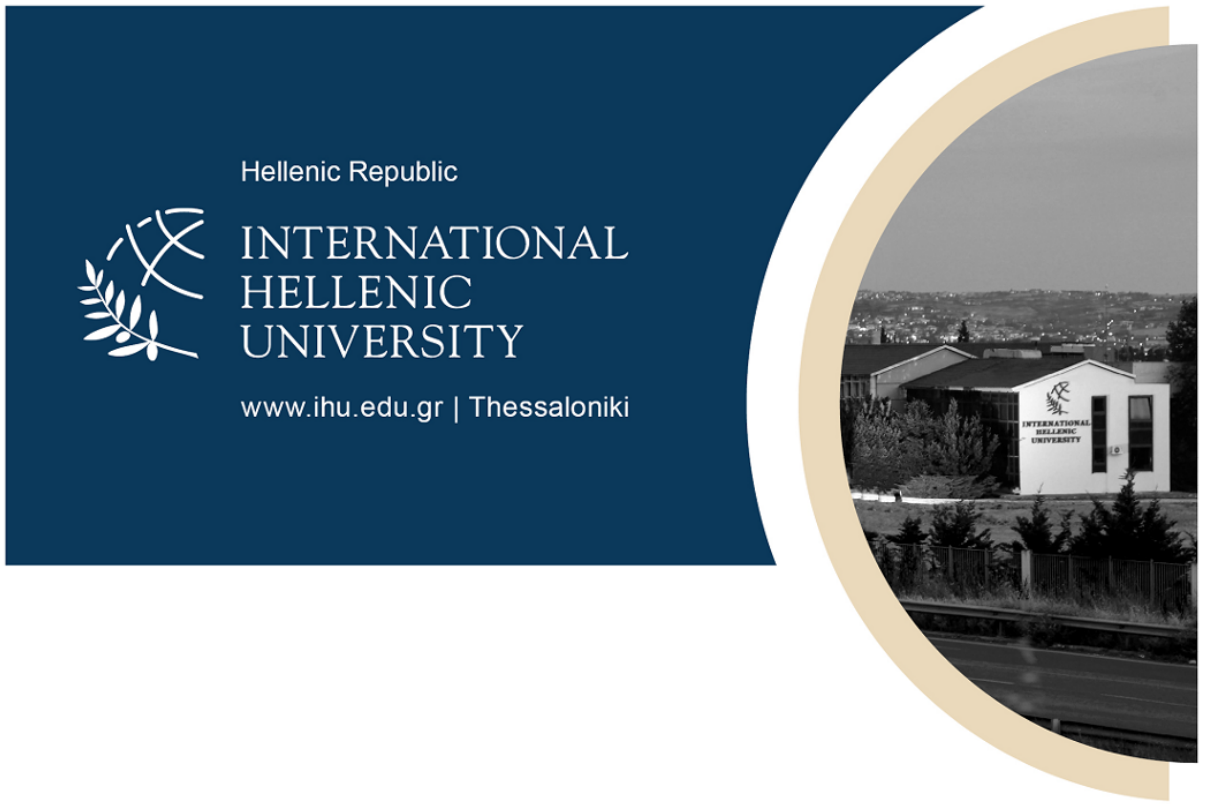
ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

Digital Organizations LAB



Ανάλυση του MITOS με χρήση τεχνικών Επεξεργασίας Φυσικής Γλώσσας και Γράφων Γνώσεις

έκδοση 1.0

Ιστορικό Εγγράφου

Ο παρακάτω πίνακας παρέχει μια επισκόπηση των πιο σημαντικών αλλαγών του εγγράφου.

Ημερομηνία	Έκδοση	Τροποποίηση	Συγγραφείς
31/10/2023	1.0	Αρχική έκδοση	Κωνσταντινίδης Ιωάννης, Μιχαηλίδης Αλέξιος, Μαγνήσαλης Ιωάννης, Μπερμπερίδης Χρήστος, Περιστέρας Βασίλειος

Πίνακας Περιεχομένων

Κατάλογος Πινάκων, Διαγραμμάτων, Εικόνων & Γραφημάτων

Κατάλογος Εικόνων

Επιτελική Σύνοψη

Στον σημερινό κόσμο που βασίζεται στα δεδομένα, η ικανότητα εξαγωγής ουσιαστικών πληροφοριών είναι πρωταρχικής σημασίας. Οι παραδοσιακές σχεσιακές βάσεις δεδομένων αποτελούν τη ραχοκοκαλιά της αποθήκευσης και ανάκτησης δεδομένων για δεκαετίες. Αποθηκεύουν δεδομένα σε δομημένους πίνακες με προκαθορισμένα σχήματα, καθιστώντας τα αποτελεσματικά για συναλλαγές. Παράλληλα, σημαντική πληροφορία αποθηκεύεται σε μορφή αδόμητου κειμένου και περιγραφών καθιστώντας την μη (άμεσα) αξιοποιήσιμη για περαιτέρω ανάλυση. Με την ραγδαία ανάπτυξη τεχνολογιών τεχνητής νοημοσύνης και προηγμένων βάσεων δεδομένων τα τελευταία χρόνια, δίνεται η δυνατότητα να τιθασεύσουμε τον μεγάλο όγκο πληροφοριών σε μια μορφή επεξεργάσιμη και αναλύσιμη από τις μηχανές. Οι γράφοι γνώσεως (knowledge graphs) αποτελούν ένα είδος σύγχρονων βάσεων δεδομένων που αναπαριστούν την πληροφορία σε μορφή κόμβων και σχέσεων και παρέχουν τη δυνατότητα εξαγωγής χρήσιμων γνώσεων μέσα από ανάλυση πολύπλοκων σχέσεων στο γράφο.

Το παρόν κείμενο έχει ως στόχο να παρουσιάσει την υλοποίηση για την ανάλυση και την επεξεργασία των δεδομένων που παρέχονται από το Εθνικό Μητρώο Διοικητικών Διαδικασιών (ΕΜΔΔ) με σκοπό να δημιουργηθεί ένας πλούσιος γράφος γνώσης με πολλές συνδέσεις μεταξύ των δεδομένων και να πραγματοποιηθεί προηγμένη ανάλυση του γράφου για εύρεση κρυμμένης πληροφορίας ως προς τις διαδικασίες. Επίσης, παρουσιάζονται η τελική αρχιτεκτονική, το διαδραστικό dashboard καθώς και η συνολική εκτίμηση της προστιθέμενης αξίας που μπορούν να επιφέρουν οι γράφοι γνώσης και η ανάλυση που πραγματοποιήθηκε στα πλαίσια του συγκεκριμένου έργου για την περαιτέρω βελτίωση και ανάπτυξης του ΕΜΔΔ.

Η μελέτη ξεκινά με την παρουσίαση του πλάνου υλοποίησης των εργαλείων και της ροής πο. Στη συνέχεια, παρουσιάζεται η τελική αρχιτεκτονική και αναλύεται το NeoDash που αποτελεί την βάση πάνω στην οποία κατασκευάστηκε το διαδραστικό dashboard, το οποίο επακολούθως παρουσιάζεται αναλυτικά. Τέλος, παρατίθενται τα συμπεράσματα της ανάλυσης που πραγματοποιήθηκε στα πλαίσια του συγκεκριμένου έργου.

1 Εισαγωγή

Τα τελευταία χρόνια παρατηρείται έντονο ενδιαφέρον για αναβάθμιση της Δημόσιας Διοίκησης με κύριες αναφορές στην απαλλαγή από τη γραφειοκρατία, καθώς και τις χρονοβόρες και πολύπλοκες διαδικασίες. Στο πλαίσιο αυτό οι κυβερνήσεις τείνουν να αναβαθμίσουν την παροχή υπηρεσιών στους πολίτες και τις επιχειρήσεις μέσω των συστημάτων ηλεκτρονικής διακυβέρνησης επιδιώκοντας την αποτελεσματική συνεργασία και την ανταλλαγή δεδομένων μεταξύ των συναρμόδιων φορέων.

Χαρακτηριστικό αποτέλεσμα αυτού του στόχου αποτελεί το Εθνικό Μητρώο Διοικητικών Διαδικασιών (ΕΜΔΔ) – Μίτος. Το ΕΜΔΔ αποτελεί μια πλατφόρμα πληροφόρησης ως προς τις διαδικασίες, τις προϋποθέσεις, τα βήματα και τα δικαιολογητικά που απαιτούνται για την διεκπεραίωσή τους καθώς και άλλες πληροφορίες. Παράλληλα, το ΕΜΔΔ παρέχεται σε μορφή ανοιχτών δεδομένων σε μορφή API (Application Programming Interface) δίνοντας τη δυνατότητα υλοποίησης εφαρμογών ανάλυσης.

Παρόλα αυτά, σημαντική γνώση για τις διαδικασίες κρύβεται σε μορφή αδόμητου κειμένου και περιγραφών καθιστώντας σύνθετη την αυτόματη αξιοποίηση από μηχανές. Με τη ραγδαία ανάπτυξη τεχνολογιών τεχνητής νοημοσύνης και επεξεργασίας φυσικής γλώσσας και της αναπαράστασή τους σε μορφή προηγμένων βάσεων δεδομένων, όπως γράφοι γνώσεως (knowledge graphs), δίνεται η δυνατότητα αυτόματης εξαγωγής δομημένης πληροφορίας από τα κείμενα καθώς και σχέσεων μεταξύ των διαδικασιών που μπορούν να αποτελέσουν τα σωστά θεμέλια για την προηγμένη ανάλυση του ΕΜΔΔ και εύρεση συμπερασμάτων που θα μπορούσαν να βοηθήσουν, για παράδειγμα, σε ανασχεδιασμό των διαδικασιών αλλά και της γενικής βελτίωσης της ποιότητας των διαδικασιών.

Οι γράφοι γνώσης αναπαριστούν δεδομένα ως κόμβους (nodes) και ακμές (relationships). Σε αντίθεση με τις σχεσιακές βάσεις δεδομένων, οι οποίες επικεντρώνονται στην αποθήκευση δεδομένων σε πίνακες, τα γραφήματα γνώσης δίνουν έμφαση στις σχέσεις μεταξύ των σημείων δεδομένων. Αυτός ο διασυνδεδεμένος ιστός πληροφοριών παρέχει μια πιο ολιστική εικόνα των δεδομένων, διευκολύνοντας την ανακάλυψη κρυμμένων μοτίβων, σχέσεων και πληροφοριών.

Ένα από τα κύρια πλεονεκτήματα των γραφημάτων γνώσης είναι η ικανότητά τους να παρέχουν το περιεχόμενο της πληροφορίας που μελετάται. Σε μια παραδοσιακή βάση δεδομένων, δεν εμφανίζονται άμεσα όλες οι συνδέσεις από τα δεδομένα καθιστώντας δύσκολο να εξαχθεί η πλήρης εικόνα. Οι γράφοι γνώσης, από την άλλη πλευρά, συνδέουν σχετικά σημεία δεδομένων, παρέχοντας μια ολοκληρωμένη άποψη του συνολικού τοπίου των δεδομένων. Για να γίνει πιο κατανοητό, ένα αντιπροσωπευτικό παράδειγμα της χρήσης γράφων γνώσης αποτελούν τα συστήματα συστάσεων (recommendation systems), όπου η κατανόηση των σχέσεων μεταξύ προϊόντων και χρηστών μπορεί να οδηγήσει σε πιο εξατομικευμένες και αποτελεσματικές προτάσεις.

Επιπλέον, οι γράφοι γνώσης είναι εγγενώς πιο ευέλικτοι από τις σχεσιακές βάσεις δεδομένων. Μπορούν εύκολα να ενσωματώσουν νέες πηγές δεδομένων, να προσαρμοστούν στις μεταβολές στις δομές και να εξελιχθούν με την πάροδο του χρόνου. Αυτή η ευελιξία είναι ιδιαίτερα σημαντική στον σημερινό ψηφιακό κόσμο, όπου τα δεδομένα αλλάζουν και αυξάνονται συνεχώς.

Συμπερασματικά, ενώ οι παραδοσιακές σχεσιακές βάσεις δεδομένων εξακολουθούν να είναι απαραίτητες για πολλές εφαρμογές, οι γράφοι γνώσης προσφέρουν έναν πιο δυναμικό και διασυνδεδεμένο τρόπο προβολής και ανάλυσης δεδομένων. Παρέχουν το πλαίσιο, την ευελιξία και το βάθος που απαιτούνται για την εξαγωγή ουσιαστικών γνώσεων από πολύπλοκα σύνολα

δεδομένων. Καθώς οι επιχειρήσεις και οι οργανισμοί συνεχίζουν να βασίζονται σε δεδομένα για τη λήψη αποφάσεων, η ανάγκη για εργαλεία όπως οι γράφοι γνώσης, που μπορούν να παρέχουν μια βαθύτερη κατανόηση των δεδομένων, θα αυξηθεί.

1.1 Σκοπός του έργου

Το παρόν κείμενο πραγματεύεται την τελική παρουσίαση της ανάλυσης της πληροφορίας του Εθνικού Μητρώου Διοικητικών Διαδικασιών (ΕΜΔΔ) όπως αυτή διατίθεται από το API του φορέα, μέσω όμως μίας διαφορετικής σκοπιάς. Πιο συγκεκριμένα, σκοπός του έργου είναι η δημιουργία ενός γράφου γνώσης πλήρως συνδεδεμένου με σκοπό την ανάδειξη κρυφών σχέσεων μεταξύ των δεδομένων που επιστρέφει το MITOS API, αλλά και επιπλέον πληροφορίας που εξάγεται μέσω επεξεργασίας αυτών με μηχανική μάθηση, η ανάλυση αυτού και η παρουσίαση του διαδραστικού dashboard:

Οι επιμέρους στόχοι του παραδοτέου είναι οι εξής:

- Παρουσίαση των εργαλείων και της ροής εργασιών που ακολουθήθηκε για την δημιουργία του γράφου γνώσης.
- Παρουσίαση και ανάλυση των μεθόδων που χρησιμοποιήθηκαν για την συλλογή και την περαιτέρω επεξεργασία των αρχικών δεδομένων.
- Εμβάθυνση στην δομή και τα στατιστικά του γράφου γνώσης που κατασκευάστηκε.
- Παρουσίαση του εργαλείου οπτικοποίησης NeoDash.
- Παρουσίαση του διαδραστικού dashboard που πραγματοποιήθηκε.
- Καταγραφή παρατηρήσεων, προκλήσεων και ενεργειών για το μέλλον, όπως αυτά προέκυψαν κατά την ανάλυση και αξιολόγηση της πληροφορίας του ΜΙΤΟΣ στα πλαίσια του πακέτου εργασίας 5.

1.2 Οργάνωση του κειμένου

Στο Κεφάλαιο 2 παρουσιάζεται το πλάνο υλοποίησης, όπως αυτό συμφωνήθηκε ύστερα από την ανάλυση απαιτήσεων και σε συνεργασία με τους ειδικούς του ΕΜΔΔ, τα εργαλεία που χρησιμοποιήθηκαν, καθώς και η ροή εργασιών που ακολουθήθηκε για την κατασκευή του γράφου γνώσης

Στο Κεφάλαιο 3 παρατίθενται και αναλύονται πληροφορίες σχετικές με τη συλλογή, προεπεξεργασία και ανάλυση των δεδομένων του Εθνικού Μητρώου Διοικητικών Διαδικασιών (ΕΜΔΔ).

Στο Κεφάλαιο 4 παρουσιάζεται η ροή εργασίας που ακολουθήθηκε για την κατασκευή του γράφου, καθώς η δομή και τα στατιστικά του.

Στο Κεφάλαιο 5, παρουσιάζεται η τελική αρχιτεκτονική και το εργαλείο οπτικοποίησης NeoDash καθώς και οι ενέργειες που απαιτούνται για την χρησιμοποίησή του.

Στο Κεφάλαιο 6, παρουσιάζονται τα αποτελέσματα του διαδραστικού dashboard.

Τέλος, στο Κεφάλαιο 7, συνοψίζονται τα κυριότερα ευρήματα της συγκεκριμένης μελέτης και προτείνονται πιθανές κατευθύνσεις σε ερευνητικό και επιχειρησιακό επίπεδο που μπορούν να συμβάλλουν στην πρόοδο του ΕΜΔΔ.

2 Πλάνο Υλοποίησης και Ροή Εργασιών

Στην παρούσα ενότητα αναλύεται το πλάνο υλοποίησης του έργου, όπως αυτό συνδιαμορφώθηκε με του ειδικούς από το ΜΔΔ στα πλαίσια της ανάλυσης απαιτήσεων. Επίσης, παρουσιάζονται τα εργαλεία που χρησιμοποιήθηκαν, καθώς και η βασική ροή εργασιών που δημιουργήθηκε για την συλλογή και επεξεργασία των δεδομένων που απαιτούνται για την κατασκευή του γράφου.

2.1 Πλάνο Υλοποίησης

Όπως παρουσιάζεται στην εικόνα 1, το πλάνο υλοποίησης διαχωρίζεται σε τέσσερα βασικά στάδια. Το πρώτο στάδιο αφορά τη Συλλογή Πληροφορίας (Information Gathering) που απαιτείται για την κατασκευή του γράφου. Πιο συγκεκριμένα, οι ενέργειες διαχωρίζονται σε τέσσερες βασικές υποκατηγορίες οι οποίες είναι η εξής: α) Συλλογή Δεδομένων από το API με βάση τα οποία συνεχίζεται η συλλογή πληροφορίας με την β) Εξαγωγή Επιπλέον Πληροφορίας (Information Extraction), γ) Δημιουργία Πρόσθετης Πληροφορίας (Information Generation) και δ) Εξαγωγή Συσχετίσεων μεταξύ των δεδομένων. Περαιτέρω στοιχεία για την συλλογή πληροφορίας παρουσιάζεται στο κεφάλαιο 3.

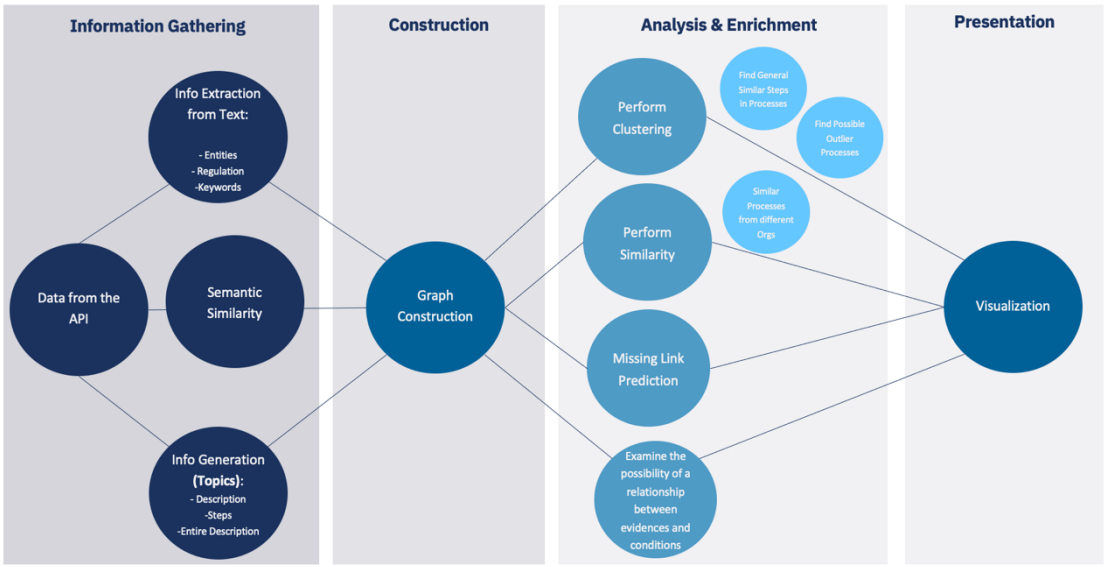
Το δεύτερο στάδιο αφορά την κατασκευή γράφου. Εφόσον έχουν συλλεχθεί τα δεδομένα, πραγματοποιείται τελική επεξεργασία (post-processing δεδομένων) ώστε να μετατραπούν τα δεδομένα σε μορφή σχέσεων και να κατασκευαστεί ο γράφος με όλη την πληροφορία που απαιτείται για να πραγματοποιηθεί η ανάλυση των use-cases που έχουν αναγνωριστεί. Περαιτέρω στοιχεία για την κατασκευή του γράφου παρουσιάζεται στο κεφάλαιο 4.

Εφόσον έχει κατασκευαστεί ο γράφος περνάμε στο επόμενο στάδιο, το οποίο αφορά την ανάλυσή του μέσω από πιθανά use-cases που έχουν εντοπιστεί. Ενδεικτικά, αυτά διαμορφώνονται στα εξής:

- Δημιουργία συμπλεγμάτων (cluster) διαδικασιών με πιθανή αναγνώριση κοινών βημάτων μεταξύ αυτών ή παράταιρων (outlier) διαδικασιών.

- Ανακάλυψη όμοιων (similarity) διαδικασιών μεταξύ τους.
- Ανακάλυψη διαδικασιών που θα έπρεπε να είναι συνδεδεμένες, αλλά δεν είναι (missing link prediction).
- Ανακάλυψη σχέσεων μεταξύ των Δικαιολογητικών και των Προϋποθέσεων μίας διαδικασίας.

Τέλος, τελευταίο στάδιο του πλάνου υλοποίησης είναι η δημιουργία διαδραστικού dashboard οπτικοποίησης, το οποίο θα προσφέρει την δυνατότητα παρουσίασης των αναλύσεων και του γράφου διαδραστικά. Αυτό το στάδιο αποτελεί παρουσιάζεται στο κεφάλαιο 6.



Εικόνα 1: Πλάνο Υλοποίησης Έργου

2.2 Εργαλεία Υλοποίησης

Για την εξόρυξη των δεδομένων και την κατασκευή του γράφου έγινε χρήση των παρακάτω εργαλείων:

KNIME[1]

[1] KNIME

Το KNIME Analytics Platform είναι ένα εργαλείο ανοιχτού κώδικα για την ανάλυση και την επεξεργασία δεδομένων. Επιτρέπει στους χρήστες να συνδέονται, να μετασχηματίζουν και να αναλύουν δεδομένα από διάφορες πηγές με σκοπό την ανακάλυψη πληροφοριών και την εξαγωγή τους σε δομημένη μορφή. Το KNIME παρέχει ένα απλό γραφικό περιβάλλον για την κατασκευή εργασιακών ροών (workflow) όπου οι χρήστες μπορούν να συνδυάζουν διάφορα μοντέλα ανάλυσης δεδομένων και μεθόδους επεξεργασίας σε σειρά. Το εργαλείο επιλέχθηκε λόγω της ευκολίας χρήσης του καθώς και της δυνατότητας αποσφαλμάτωσης (debugging) και εύκολης επαναχρησιμοποίησης από διαφορετικούς χρήστες. Παράλληλα, το KNIME επιτρέπει την ενσωμάτωση διάφορων τεχνολογιών (όπως Neo4j) άλλα και γλώσσεων προγραμματισμού όπως Python, όπου μπορούμε να χρησιμοποιήσουμε και πιο σύνθετες βιβλιοθήκες όπου χρειάζεται (πχ information extraction from text).

Επομένως, δίνει την δυνατότητα εξαγωγής των δεδομένων, π.χ από το MITOS API (Ενότητα 3), μετατροπής σε πιο ευέλικτη μορφή, όπως πινάκων και ανάλυσης αυτών με την ενσωμάτωση τεχνολογιών γράφου (πχ Neo4j) που είναι στα πλαίσια του σκοπού του παραδοτέου.

SharePoint OnLine [2]

Το SharePoint είναι μια πλατφόρμα συνεργασίας και ενσωματώνεται στο Microsoft Office. Έχει υψηλές δυνατότητες παραμετροποίησης όμως συνηθώς αποτελεί πλατφόρμα για την διαχείριση και αποθήκευση αρχείων σε οργανισμούς και επιχειρήσεις. Μέρος του αποτελεί το SharePoint OnLine που συνήθως είναι άμεσα συνδεδεμένο με το Microsoft 365. Ένα από τα πλεονέκτημα του είναι ότι δεν χρειάζεται εγκατάσταση και συντήρηση σε υποδομές του οργανισμού ή της επιχείρησης που το χρησιμοποιεί καθώς όλη η πλατφόρμα παρέχεται ως “Πλατφόρμα ως υπηρεσία” (Platform as a Service-PaaS).

[2] Microsoft 365 - Συνδρομή για εφαρμογές του Office | Microsoft 365

Python[3]

Η Python είναι αντικειμενοστραφής, γενικού σκοπού γλώσσα προγραμματισμού που χαρακτηρίζεται από αναγνωσιμότητα και ευκολία στη χρήση. Υποστηρίζει διάφορα πακέτα που επιτρέπουν την επαναχρησιμοποίηση του κώδικα καθώς επίσης και πολλές βιβλιοθήκες και εργαλεία που διευκολύνουν την ανάλυση των δεδομένων. Στην περίπτωση μας χρησιμοποιείται τόσο εντός των ροών εργασιών του KNIME, όσο και στο Google Colab, το οποίο χρησιμοποιείται καθώς απαιτείται GPU και γι' αυτό τον λόγο δεν μπορούμε να ενσωματώσουμε τις αναλύσεις σε μία ροή εργασιών.

Neo4j[4]

Το Neo4j είναι ένα σύστημα διαχείρισης βάσεων δεδομένων βασισμένο στο μοντέλο γράφων. Αποτελεί μια πλατφόρμα που επιτρέπει την αποθήκευση, την επεξεργασία και την ανάκτηση δεδομένων με τρόπο που αντιπροσωπεύει τις σχέσεις μεταξύ των δεδομένων. Σε αντίθεση με τις παραδοσιακές βάσεις δεδομένων που βασίζονται σε πίνακες, το Neo4j αποτελεί ένα είδος property graph, όπου ο γράφος απαρτίζεται από κόμβους (nodes) που αφορούν μια οντότητα (πχ διαδικασία, βήματα διαδικασίας, οργανισμός), οι οποίοι περιλαμβάνουν κάποιες ιδιότητες (πχ όνομα, id και ό,τι άλλο οριστεί). Παράλληλα, οι κόμβοι συνδέονται μεταξύ τους με σχέσεις (relationships, πχ ένας συγκεκριμένος κόμβος τύπου Διαδικασία → HAS_EVIDENCE → ένα συγκεκριμένο κόμβο τύπου EVIDENCE), δημιουργώντας ένα πλούσιο γράφο γνώσης. Μερικά σημαντικά στοιχεία για το Neo4j συνοψίζονται στα εξής:

[3] Python
[4] Neo4j

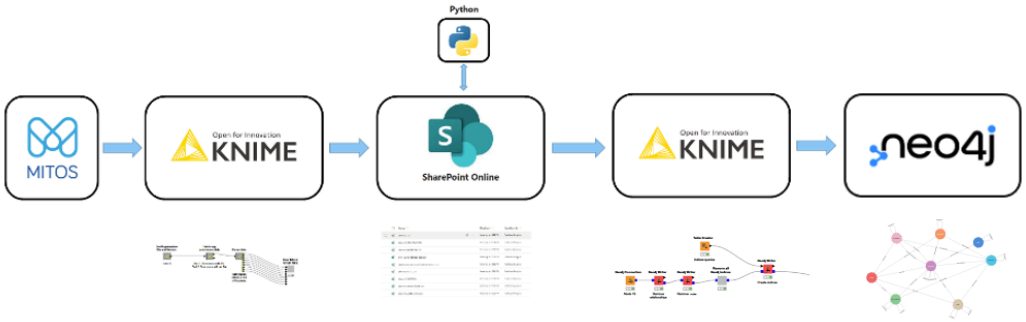
- Cypher Query Language: Το Neo4j χρησιμοποιεί τη δική του γλώσσα ερωτημάτων που ονομάζεται Cypher, η οποία έχει σχεδιαστεί ειδικά για την αναζήτηση δεδομένων γραφήματος.
- Απόδοση: Το Neo4j είναι βελτιστοποιημένο για γρήγορες διελεύσεις γραφημάτων, καθιστώντας το αποτελεσματικό για λειτουργίες που είναι προκλητικές σε παραδοσιακές σχεσιακές βάσεις δεδομένων.
- Ευελιξία: Οι βάσεις δεδομένων γραφημάτων όπως το Neo4j δεν έχουν σχήματα, πράγμα που σημαίνει ότι μπορείτε να προσθέσετε νέους τύπους δεδομένων χωρίς να χρειάζεται να τροποποιήσετε ένα υπάρχον σχήμα.
- Ενσωμάτωση: Το Neo4j μπορεί να ενσωματωθεί με διάφορα εργαλεία και πλατφόρμες και προσφέρει προγράμματα οδήγησης για πολλές γλώσσες προγραμματισμού.
- Κοινότητα και εμπορικές εκδόσεις: Το Neo4j είναι διαθέσιμο τόσο σε έκδοση κοινότητας, η οποία είναι ανοιχτού κώδικα, όσο και σε εμπορική έκδοση με πρόσθετες δυνατότητες και υποστήριξη.

2.3 Ροή Εργασιών

Στην ενότητα αυτή περιγράφονται οι βασικές ενέργειες που απαιτούνται για την ανάλυση των δεδομένων του ΕΜΔΔ, από τη συλλογή έως την κατασκευή του γράφου (Εικόνα 2).

Η ροή εργασίας υλοποιήθηκε έτσι ώστε να γίνεται η εξόρυξη των δεδομένων από το Mitos API με την χρήση του KNIME. Στην συνέχεια, χρησιμοποιήθηκε το SharePoint ως ένα ενδιάμεσο στάδιο για την αποθήκευση των δεδομένων σε μορφή CSV ώστε να μπορούν τα δεδομένα από το Mitos να αξιοποιηθούν και από άλλα εργαλεία. Τα δεδομένα αυτά επεξεργάστηκαν με την χρήση KNIME και python για το στάδιο της εξαγωγής περαιτέρω πληροφορίας και σχέσεων, τα οποία αποθηκεύτηκαν στο SharePoint. Παράλληλα, δημιουργήθηκε η βάση δεδομένων γράφου στο Neo4j. Τέλος, μέσω μιας ροής εργασιών στο KNIME

πραγματοποιείται ο τελικός καθαρισμός των δεδομένων και στην συνέχεια καταχωρούνται η μορφή και τα δεδομένα του γράφου.



Εικόνα 2: Ροή Εργασιών

Πιο αναλυτικά, ακολουθήθηκαν τα παρακάτω βήματα:

I. Ανάλυση του Mitos API ώστε να κατανοηθεί η δομή στην οποία επιστρέφει τα δεδομένα για τις διαδικασίες.
II. Χρήση του KNIME για την εξόρυξη της πληροφορίας διαδικασιών από το Mitos API.
III. Επεξεργασία και διαμόρφωση των δεδομένων εντός του KNIME ώστε να μπορούν να αποθηκευτούν σε μορφή .CSV.
IV. Αποθήκευση των .CSV αρχείων στο SharePoint OnLine.
V. Εξαγωγή των λέξεων/φράσεων κλειδιών μέσω KNIME και αποθήκευση σε .CSV μορφή στο SharePoint OnLine.
VI. Εξαγωγή επιπλέον πληροφορίων με την χρήση Python στο Google Colab και αποθήκευση σε .CSV μορφή στο SharePoint OnLine.
VII. Δημιουργία κενής βάσης δεδομένων στο Neo4j.
VIII. Άντληση των αρχείων με το KNIME.
IX. Καθαρισμός των δεδομένων και εξαγωγή περαιτέρω πληροφορίας μέσω της Ροής εργασιών του KNIME.

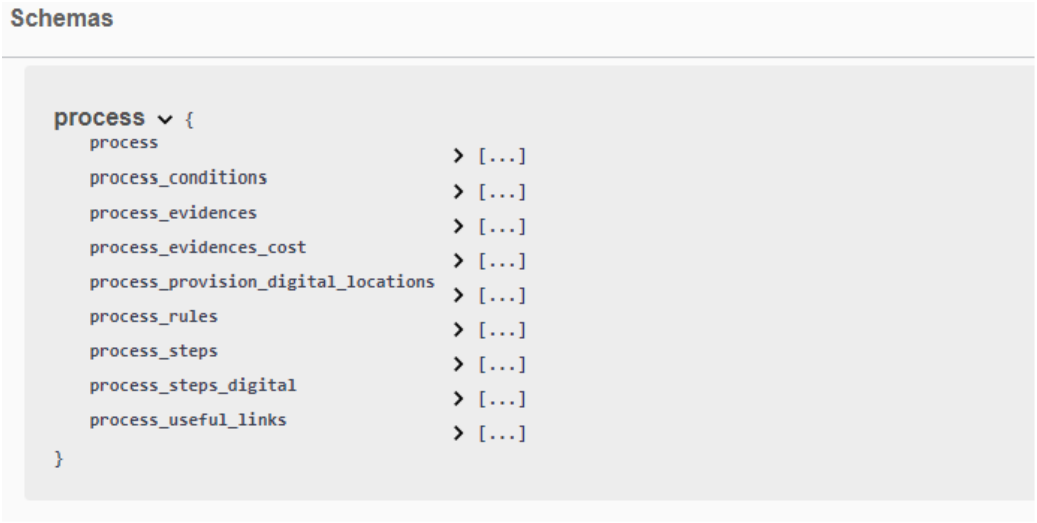
3 Συλλογή και Επεξεργασία Δεδομένων

Το ΕΜΔΔ είναι το επίσημο μητρώο όλων των διαδικασιών του ελληνικού δημοσίου. Το πληροφοριακό αυτό σύστημα αποσκοπεί στη δημιουργία ενός μητρώου για τη διαχείριση των διαδικασιών των δημοσίων υπηρεσιών και φορέων. Το μητρώο απευθύνεται σε ιδιώτες, επαγγελματίες και στελέχη της δημόσιας διοίκησης, που ενδιαφέρονται να πληροφορηθούν για τον τρόπο, χρόνο και κόστος υλοποίησης των διαδικασιών του δημόσιου τομέα.

Υπάρχουν 4 τύποι διαδικασιών που μπορούν να αναζητηθούν από το χρήστη: δημοσιευμένες, προς δημοσίευση, προς έγκριση, υπό επεξεργασία.

Ένα βασικό πλεονέκτημα του ΕΜΔΔ είναι η παροχή API ανοιχτών δεδομένων (open data) που επιτρέπει την εύκολη πρόσβαση και επεξεργασία του φάσματος των διαδικασιών. Το API (Application Programming Interface ή Διεπαφή Προγραμματισμού Εφαρμογών) Mitos προσφέρει την δυνατότητα άντλησης δεδομένων σχετικά για τις διαδικασίες όπως αυτές περιγράφονται στο Εθνικό Μητρώο Διοικητικών Διαδικασιών. Το Mitos API ακολουθεί την αρχιτεκτονική REST (Representational State Transfer) σύμφωνα με το openAPI 3.0. Η χρήση του Mitos API γίνεται με την χρήση αιτημάτων HTTP (Hypertext Transfer Protocol).

Τα δεδομένα των διαδικασιών επιστρέφονται από το Mitos API με μορφή JSON (Java Script Object Notation) και έχει την ακόλουθη δομή όπως φαίνεται στην παρακάτω εικόνα.



Εικόνα 3: Σχήμα απάντησης από το Mitos API για μια διαδικασία

Το σχήμα αναπτύσσεται επιπλέον για όλα τα μέρη που αποτυπώνονται στην παραπάνω εικόνα όπως process, process_evidences, process_evidences_cost κλπ. Η παρακάτω εικόνα παρουσιάζει την ανάπτυξη για τα process_evidences όπως αυτά δομούνται σύμφωνα με το Mitos API

Εικόνα 4: Σχήμα απάντησης δικαιολογητικών από το Mitos API για μια διαδικασία

Σε πρώτο στάδιο αντλείται η πληροφορία από το Mitos API (JSON) μέσω του KNIME[5] και μετατρέπεται σε μορφή CSV, ώστε να γίνεται ευκολότερη η ανάλυση και παράλληλα να μπορεί να αξιοποιείται και από άλλα εργαλεία ανάλυσης όπως στο πακέτα εργασίας 1 με τη χρήση του εργαλείου Power BI.

3.1 Information Extraction

Παρακάτω, αναλύονται οι τεχνικές που χρησιμοποιήθηκαν και η πληροφορία που εξήχθησε στο πλαίσιο της συλλογής επιπλέον πληροφορίας για τον εμπλουτισμό του γράφου.

3.1.1 Keyword Extraction

Η πρώτη διεργασία εξόρυξης πληροφορίας από τα δεδομένα του Mitos API, αφορά την εξαγωγή λέξεων/φράσεων κλειδιών στο κείμενο που υπάρχει. Πιο συγκεκριμένα, εξήχθησε πληροφορία από τους τίτλους των διαδικασιών, των βημάτων, των ψηφιακών βημάτων, των προϋποθέσεων, των δικαιολογητικών και των κανόνων.

Για τα πραγματοποιηθεί αυτό δημιουργείται μία ροή εργασιών στο KNIME[6] (εικονα 5), όπου διαβάζει αρχικά τα CSV αρχεία που είναι αποθηκευμένα στο SharePoint. Στην συνέχεια, χρησιμοποιείται η βιβλιοθήκη KeyBERT[7], η οποία εκμεταλλευόμενη τη μετατροπή του κειμένου σε embeddings, μέσω του προ-εκπαιδευμένου πολυγλωσσικού μοντέλου paraphrase-multilingual-mpnet-base-v2[8] (το μοντέλο επιλέχθηκε ως το πιο αποτελεσματικό πολυγλωσσικό μοντέλο για χρήση σημασιολογικής ομοιότητας σύμφωνα με την υφιστάμενη ανάλυση[9]), έχει

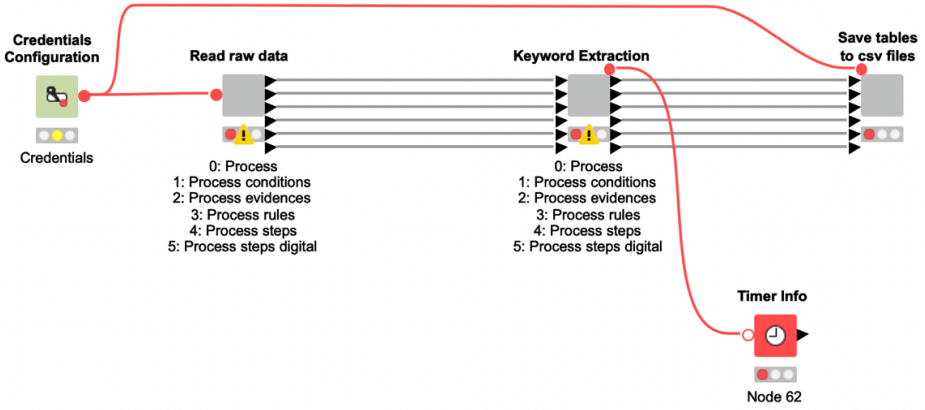
[5] KNIME Workflow
[6] Keyword Extraction Pipeline
[7] https://github.com/MaartenGr/KeyBERT
[8] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2
[9] https://www.sbert.net/docs/pretrained_models.html

την δυνατότητα αναγνώρισης βασικών λέξεων - φράσεων κλειδιών στις προτάσεις. Οι φράσεις που εξάγονται αποθηκεύονται σε μορφή CSV στο SharePoint για κάθε αρχείο ξεχωριστά.

Ακολούθως, οι λέξεις – φράσεις κλειδιά χρειάζεται να ομογενοποιηθούν (keyword disambiguation) και να αποκτήσουν ένα μοναδικό ID. Για να γίνει αυτό, η ανάλυση γίνεται στο Google Colab για ταχύτερη εκτέλεση μέσω χρήσης GPU, όπου πραγματοποιείται ομαδοποίηση (clustering) των keywords[10]. με την χρήση του μοντέλου BERTopic[11] σε κάθε αρχείο (process, process_evidence, process_condition, process_step, process_digital_step), και στην συνέχεια πραγματοποιείται έλεγχος σημασιολογικής ομοιότητας[12]. των λέξεων - φράσεων μέσα στο ίδιο cluster με χρήση BERT embeddings και την μαθηματική φόρμουλα cosine similarity[13], και αντικατάσταση σε όσα έχουν ομοιότητα πάνω από 93% με μία **κοινή λέξη – φράση**. Τέλος, πραγματοποιείται **lemmatization.**

Εφόσον οι λέξεις – φράσεις βρίσκονται πλέον στην κατάλληλη μορφή, δημιουργείται ένα αρχείο με όλα τα keywords, μέσω του οποίου εξάγεται ένα ID για κάθε μοναδική λέξη - φράση. Επίσης, κάθε φράση συνδέεται και με τα IDs των λέξεων που έχει, πληροφορία η οποία αποθηκεύεται στο SharePoint ως CSV.

[10] Keywords Clustering Notebook
[11] https://maartengr.github.io/BERTopic/index.html
[12] Keywords Disambiguation Notebook
[13] http://singhal.info/ieee2001.pdf



Εικόνα 5: Ροή Εργασίας στο KNIME για εξαγωγή Λέξεων – Φράσεων κλειδιών

3.1.2 Entities Extraction

Όσον αφορά την εξαγωγή οργανισμών (entities) από το κείμενο, αρχικά χρησιμοποιείται μία fine-tuned version του Greek Bert[14], που έχει εκπαιδευτεί για να αναγνωρίζει οργανισμούς στα Ελληνικά, καθώς και το pipeline που παρέχει το HuggingFace για Named Entity Recognition[15]. Όπως και προηγουμένως, αυτή η διεργασία πραγματοποιείται για κάθε αρχείο ξεχωριστά μόνο που αυτή την φορά αντί για το KNIME, χρησιμοποιείται το Google Colab με GPU για ταχύτερη εκτέλεση[16]. Εφόσον αποθηκευτούν τα αρχεία στο SharePoint, η ανάλυση που ακολουθείται είναι παρόμοια με το Keyword Extraction προηγουμένως[17] [18].

Παρόλο που η εξαγωγή οργανισμών πραγματοποιήθηκε, δεν χρησιμοποιήθηκε η πληροφορία στον γράφο, καθώς μερικοί οργανισμοί έχουν αποτυπωθεί και από το Keyword Extraction, ενώ επίσης υπάρχει παρόμοια πληροφορία μέσω του API (org-owner).

[14] https://huggingface.co/amichailidis/bert-base-greek-uncased-v1-finetuned-ner
[15] https://huggingface.co/docs/transformers/main_classes/pipelines
[16] Entities Extraction Notebook
[17] Entities Clusterting Notebook
[18] Entities Disambbiguation Notebook

3.1.3 Regulation Extraction

Η εξαγωγή νομοθεσιών που υπάρχουν μέσα στο κείμενο, όπως τα προεδρικά διατάγματα, πραγματοποιήθηκε σε κάθε αρχείο ξεχωριστά με την χρήση regex pattern[19], όπως αυτό παρουσιάζεται στην εικόνα 6. Στην συνέχεια η πληροφορία που εξήχθησε συγκεντρώθηκε σε ένα προσωρινό αρχείο, με σκοπό να δοθεί για κάθε νομοθεσία ένα μοναδικό ID. Ακολούθως σώθηκε η πληροφορία για κάθε αρχείο ξεχωριστά.

Παρόλο που η εξαγωγή νομοθεσιών πραγματοποιήθηκε, δεν χρησιμοποιήθηκε η πληροφορία στον γράφο, καθώς αυτή έχει μερικώς αποτυπωθεί και από το Keyword Extraction.

```
def extract_regulations(df, column_name, id):

    column = column_name
    id = id
    pattern = r'(\S+\s(?:\d+\/(?:\d+(?:\.\d+)*(?:\/\d+(?:-\d+)?)?)?(?:-\d+)?))'
```

Εικόνα 6: Regex Pattern for Regulation Extraction

3.2 Topic Extraction

Η εξαγωγή Topics, δηλαδή σημασιολογικά κοινών ομάδων, αποτελεί μία διεργασία δημιουργίας πρόσθετης πληροφορίας από τα δεδομένα του Mitos API με σκοπό την ανάπτυξη περισσότερων συνδέσεων μεταξύ των στοιχείων του γράφου. Ειδικότερα, τα CSV αρχεία εισάγονται στο Google Colab (χρήση GPU), διαχωρίζεται η πληροφορία που μας ενδιαφέρει (τίτλοι των διαδικασιών, των βημάτων, των ψηφιακών βημάτων, των προϋποθέσεων και των δικαιολογητικών) και ενοποιούνται όλα σε ένα προσωρινό κοινό αρχείο.

Στο συνολικό πλέον αρχείο πραγματοποιείται η διαδικασία του Topic Modeling, αξιοποιώντας την βιβλιοθήκη BERTopic και κατασκευάζοντας ένα μοντέλο με σκοπό την δημιουργία ομάδων με παρόμοια σημασιολογική έννοια (ύστερα από

[19] Regulation Extraction Notebook

μετατροπή του κειμένου σε διανύσματα με τη χρήση του προεκπαιδευμένου μοντέλου BERT paraphrase-multilingual-mpnet-base-v2). Ένας μεγάλος αριθμός των δεδομένων κατηγοριοποιείται ως outlier (Topic -1), γεγονός το οποίο προσπαθούμε να μειώσουμε. Αντίστοιχα ενοποιούμε και κάποια σημασιολογικά κοντινά Topics που δεν έχουν ενοποιηθεί από το μοντέλο μας.

Εφόσον έχουμε πλέον τα Topics μας στην κατάλληλη μορφή, εξάγουμε για κάθε αρχείο την πληροφορία του Topic που ανήκει, καθώς και ένα γενικό αρχείο με επιπλέον πληροφορία για κάθε Topic, όπως ο αριθμός των αρχείων που ανήκουν σε κάθε Topic, ένα αντιπροσωπευτικό όνομα, αντιπροσωπευτικές λέξεις και αντιπροσωπευτικά αρχεία του συγκεκριμένου Topic (εικόνα 7). Η παραπάνω διαδικασία συγκεντρώνεται σε ένα notebook[20].

	Topic	Count	Name	Representation	Representative_Docs
0	-1	2650	-1_ηλικίας_τουλάχιστον_έτος_έτη	[ηλικίας, τουλάχιστον, έτος, έτη, μόνο, άλλη, ...	[Οι υποψήφιοι οδηγοί και οδηγοί προκειμένου να...
1	0	1259	0_πλοίου_ναυτικού_πλοίο_σκάφους	[πλοίου, ναυτικού, πλοίο, σκάφους, πλοία, αλιε...	[Να είναι επιβατηγό πλοίο ή ταχύπλοο σκάφος πο...
2	1	762	1_κυκλοφορίας_οχήματος_οχημάτων_οδήγησης	[κυκλοφορίας, οχήματος, οχημάτων, οδήγησης, άδ...	[Άδεια κυκλοφορίας οχήματος σε ισχύ, του δηλού...

Εικόνα 7: Πληροφορία των Topic -1 (outlier), 0 και 1

3.3 Semantic Similarity

Τέλος, ένα πολύ σημαντικό βήμα της επεξεργασίας των δεδομένων είναι η ομογενοποίηση τους. Για παράδειγμα, η εύρεση βημάτων που είναι ίδια αλλά γράφονται διαφορετικά (π.χ Εκκίνηση Διαδικασίας & Εκκίνηση της Διαδικασίας). Σκοπός αυτής της διεργασίας είναι η σύνδεση αυτής της πληροφορίας ώστε να μπορέσουμε να έχουμε περισσότερες συνδέσεις μέσα στον γράφο. Για τον σκοπό αυτό, δημιουργείται μία ροή εργασιών με δύο βασικά βήματα.

Αρχικά, μετατρέπονται τα κείμενα σε διανύσματα (embeddings) και πραγματοποιείται clustering των δεδομένων ξεχωριστά (τίτλοι των διαδικασιών, των βημάτων, των ψηφιακών βημάτων, περιγραφές των προϋποθέσεων και των δικαιολογητικών), όπως αυτό επιγράφθηκε στην προηγούμενη ενότητα 3.2.

[20] Topic Modelling Notebook

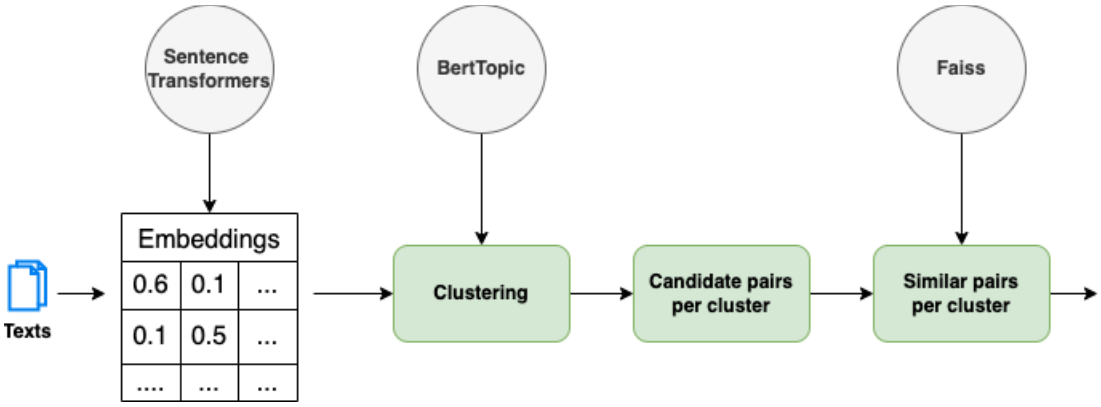
Δημιουργώντας topics (clusters)[21], μειώνουμε τα ζευγάρια που πρόκειται να συγκριθούν, καθώς πλέον οι συγκρίσεις θα πραγματοποιηθούν μόνο εντός του σημασιολογικά κοινού cluster. Αυτή η πληροφορία για κάθε αρχείο σώζεται στο SharePoint.

Στην συνέχεια και για κάθε αρχείο ξεχωριστά πραγματοποιείται ένας βρόχος (loop), όπου:

1. Δημιουργούνται Embeddings με το πολυγλωσσικό μοντέλο (**paraphrase-multilingual-mpnet-base-v2**) για τις προτάσεις του cluster που θα επιλεγεί.
2. Κάθε document (πχ τίτλος του βήματος) συγκρίνεται με μία λίστα unique πχ βημάτων που δημιουργείται στην πορεία (Greedy Approach), ενώ χρησιμοποιείται η βιβλιοθήκη FAISS[22] (Facebook AI Similarity Search) για τον αποτελεσματικότερο έλεγχο των ομοιοτήτων.
3. Τα documents συγκρίνονται αν δεν είναι όμοια πάνω από ένα όριο που ορίζουμε εμείς (στην περίπτωση μας 93%), τότε προστίθενται στην λίστα. Ειδάλλως αντικαθίστανται με το όμοιο.

Έχοντας πλέον την πληροφορία της ομοιότητας μεταξύ των documents του αρχείου και με ποιο αντικαθίστανται εφόσον έχουν ομοιότητα μεγαλύτερη από το 93% που έχουμε θέσει την σώζουμε σε μορφή CSV στο SharePoint. Οι διαδικασίες που ακολουθούνται συγκεντρώνονται σε ένα notebook[23].

[21] Topic Modelling Noteboko (towards harmonization)
[22] https://github.com/facebookresearch/faiss
[23] Semantic Similarity Notebook



Εικόνα 8 : Αρχιτεκτονική semantic similarity

Στην Εικόνα 8 αποτυπώνεται συνοπτικά η αρχιτεκτονική της μεθοδολογίας για semantic similarity.

4 Κατασκευή Γράφου

4.1 KNIME workflow

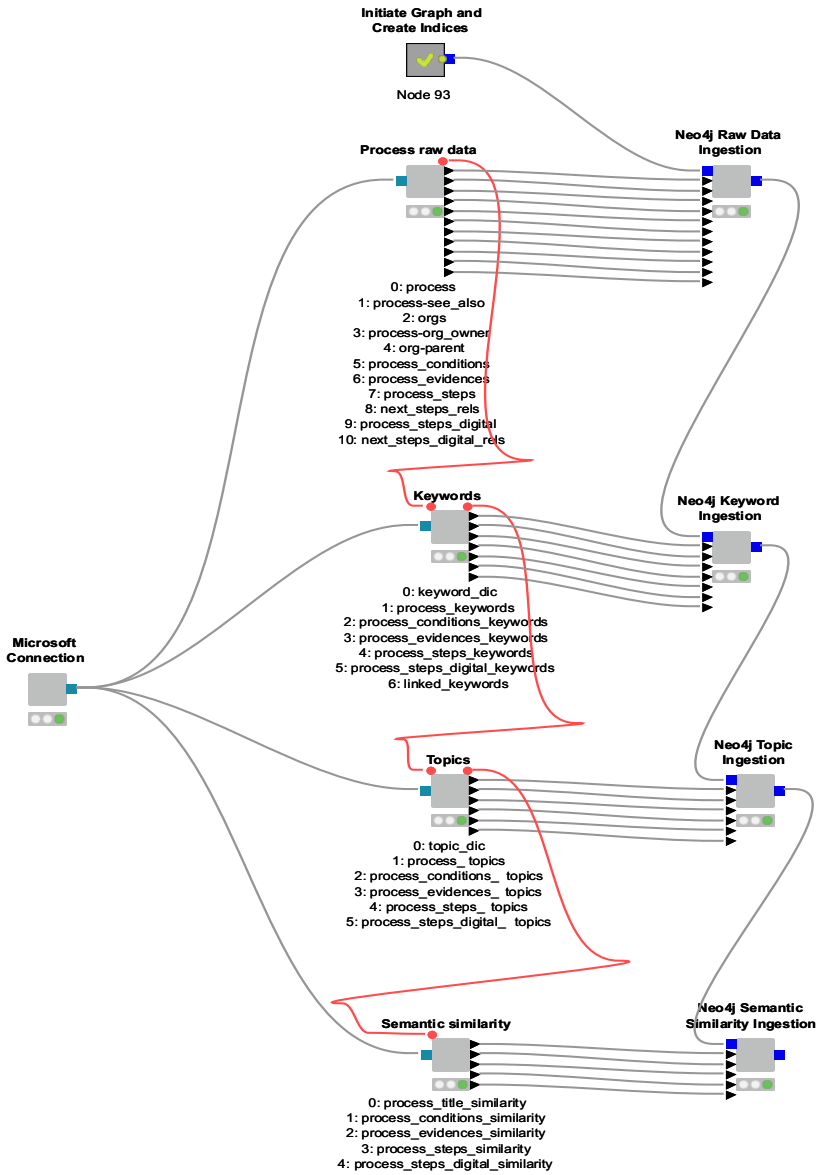
Έχοντας ολοκληρώσει τη συλλογή, επεξεργασία και εξαγωγή δεδομένων του Mitos API, υπάρχει η δυνατότητα κατασκευής του γράφου για να πραγματοποιηθεί προηγμένη ανάλυση. Για το σκοπό αυτό, δημιουργήθηκε ένα KNIME workflow[24], το οποίο συλλέγει τα επεξεργασμένα δεδομένα αποθηκευμένα σε μορφή CSV στο Microsoft SharePoint και δημιουργεί το γράφο. Πιο συγκεκριμένα, το workflow αποτελείται από τα ακόλουθα βήματα:

- **Δημιουργία δομής γράφου**: Εφόσον, έχει δημιουργηθεί ένας κενό γράφος στο Neo4j, μπορούμε να ορίσουμε τη δομή του γράφου. Συγκεκριμένα, διαγράφουμε ότι δεδομένα έχει ο γράφος σε περίπτωση επανεκτέλεσης του workflow και ορίζουμε indices για τους διαφορετικούς τύπους κόμβων (nodes) του γράφου. Αυτό αποτελεί σημαντικό παράγοντα στη διεργασία καθώς μειώνεται δραστικά ο χρόνος εκτέλεσης των queries που στέλνουμε στο γράφο.
- **Συλλογή επεξεργασμένων δεδομένων**: Σε αυτό το στάδιο, το workflow συνδέεται στο SharePoint και συλλέγει τα επεξεργασμένα δεδομένα αποθηκευμένα σε μορφή CSV. Συγκεκριμένα, συλλέγονται τα δεδομένα του MITOS με τις διαδικασίες, τα βήματα, τα ψηφιακά βήματα, τις προϋποθέσεις και τα δικαιολογητικά, καθώς και τα εξαγόμενα keywords, topics και σχέσεις σημασιολογικής ομοιότητας (semantic similarity).
- **Μετα-επεξεργασία των δεδομένων για εισαγωγή στο γράφο**: Τα δεδομένα που εισάγονται στο KNIME, επεξεργάζονται περαιτέρω ώστε να δημιουργηθούν πίνακες με τους κόμβους και τις σχέσεις μεταξύ τους ώστε να εισαχθούν στο Neo4j.

[24] KNIME workflow

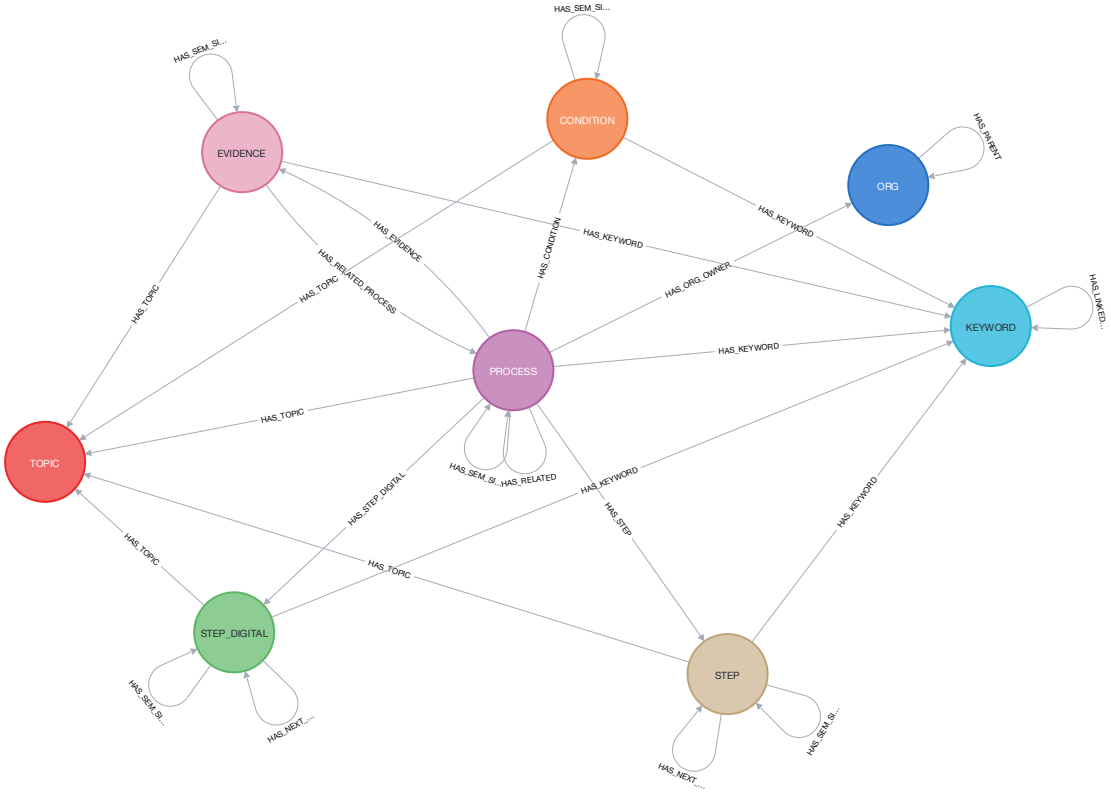
- **Εισαγωγή δεδομένων στο γράφο**: Χρησιμοποιούνται οι πίνακες με τους κόμβους και τις σχέσεις και δημιουργούνται queries στο Neo4j που εισάγουν τα δεδομένα.

Παρακάτω, απεικονίζεται το δεδομένο workflow κατασκευής του γράφου σε συνοπτική μορφή αποτελούμενο από components (sub-workflows), όπου φαίνεται και η δυνατότητα documentation και ευκολότερης κατανόησης της διεργασίας του workflow στο KNIME.



Εικόνα 9: Ροή Εργασίας στο KNIME για κατασκευή γράφου

4.2 Δομή Γράφου



Εικόνα 10: Σχήμα γράφου στο Neo4j

Στην Εικόνα 10 απεικονίζεται το σχήμα του γράφου όπου φαίνονται οι τύποι κόμβων καθώς και οι σχέσεις μεταξύ τους. Γίνεται αντιληπτό πως ύστερα από την συλλογή-επεξεργασία δεδομένων και την κατασκευή του γράφου, οι κόμβοι παρουσιάζουν πολλές συνδέσεις μεταξύ τους, δίνοντας επομένως τη δυνατότητα σύνθετης ανάλυσης γράφου σύμφωνα με τις συνδέσεις αυτές. Για παράδειγμα, μπορούμε να εντοπίσουμε συσχετίσεις μεταξύ διαδικασιών χρησιμοποιώντας συνδυαστικά πληροφορία από όλους τους κόμβους που αφορούν τη διαδικασία, αυξάνοντας την αποτελεσματικότητα της ανάλυσης. Στους Πίνακες 1 και 2 περιγράφονται, πιο συγκεκριμένα, οι κόμβοι και οι σχέσεις του γράφου, αντίστοιχα.

Πίνακας 1: Περιγραφή κόμβων (nodes) στο Neo4j

Κόμβος	Περιγραφή
PROCESS	Γενικά στοιχεία της διαδικασίας (τίτλος, D)
ORG	Οργανισμός (όνομα, D)
CONDT ON	Προϋπόθεση διαδικασίας (περιγραφή, D)
EV DENCE	Δικαιολογητικό διαδικασίας (περιγραφή, D)
STEP	Βήμα διαδικασίας (περιγραφή, D)
STEP_D G TAL	Ψηφιακό βήμα διαδικασίας (περιγραφή, D)
KEYWORD	Λέξεις κλειδιά που εξάχθηκαν από τις περιγραφές των παραπάνω
TOP C	Θεματολογίες – clusters των περιγραφών των παραπάνω

Πίνακας 2: Περιγραφή σχέσεων (relationships) στο Neo4j

Κόμβος	Περιγραφή
HAS_RELATED	Σχέση μεταξύ διαδικασιών σύμφωνα με το see_also από το M TOS
HAS_ORG_OWNER	Σχέση μεταξύ διαδικασιών και οργανισμών που δείχνει τον οργανισμό που αναλαμβάνει τη διαδικασία
HAS_PARENT	Σχέση μεταξύ οργανισμών και οργανωτικών μονάδων που δείχνει την ιεραρχία μεταξύ τους (αν υπάγεται κάποια μονάδα σε κάποιο φορέα/μονάδα)
HAS_CONDT ON	Σχέση μεταξύ διαδικασιών και προϋποθέσεων που δείχνει κάθε προϋπόθεση που αφορά μια διαδικασία
HAS_EV DENCE	Σχέση μεταξύ διαδικασιών και δικαιολογητικών που δείχνει κάθε δικαιολογητικό που απαιτεί μια διαδικασία

HAS_RELATED_PROCESS	Σχέση μεταξύ διαδικασιών και δικαιολογητικών που δείχνει ποια είναι η σχετική διαδικασία λήψης του ζητούμενου δικαιολογητικού
HAS_STEP	Σχέση μεταξύ διαδικασιών και βημάτων που δείχνει τα βήματα που χρειάζονται για μια διαδικασία
HAS_NEXT_STEP	Σχέση μεταξύ βημάτων που δείχνει το επόμενο βήμα σε μια διαδικασία
HAS_STEP_DIGITAL	Σχέση μεταξύ διαδικασιών και ψηφιακών βημάτων που δείχνει τα ψηφιακά βήματα που χρειάζονται για μια διαδικασία
HAS_NEXT_STEP_DIGITAL	Σχέση μεταξύ ψηφιακών βημάτων που δείχνει το επόμενο ψηφιακό βήμα σε μια διαδικασία
HAS_KEYWORD	Σχέση μεταξύ κόμβων (PROCESS, EVIDENCE, CONDITION, STEP, STEP_DIGITAL) και keywords που δείχνει τις σχετικές λέξεις-κλειδιά για έναν κόμβο
HAS_LINKED_KEYWORD	Σχέση μεταξύ keywords που δείχνει τις λέξεις-κλειδιά που εμπεριέχονται σε μία φράση κλειδί
HAS_TOPIC	Σχέση μεταξύ κόμβων (PROCESS, EVIDENCE, CONDITION, STEP, STEP_DIGITAL) και topics που δείχνει το θεματικό cluster που ανήκει ένας κόμβος
HAS_SEM_SIMILAR_PROCESS	Σχέση μεταξύ διαδικασιών που δείχνει την ύπαρξη σημασιολογικής ομοιότητας μεταξύ των τίτλων των διαδικασιών
HAS_SEM_SIMILAR_CONDITION	Σχέση μεταξύ προϋποθέσεων που δείχνει την ύπαρξη σημασιολογικής ομοιότητας μεταξύ των περιγραφών των προϋποθέσεων
HAS_SEM_SIMILAR_EVIDENCE	Σχέση μεταξύ δικαιολογητικών που δείχνει την ύπαρξη σημασιολογικής ομοιότητας μεταξύ των περιγραφών των δικαιολογητικών

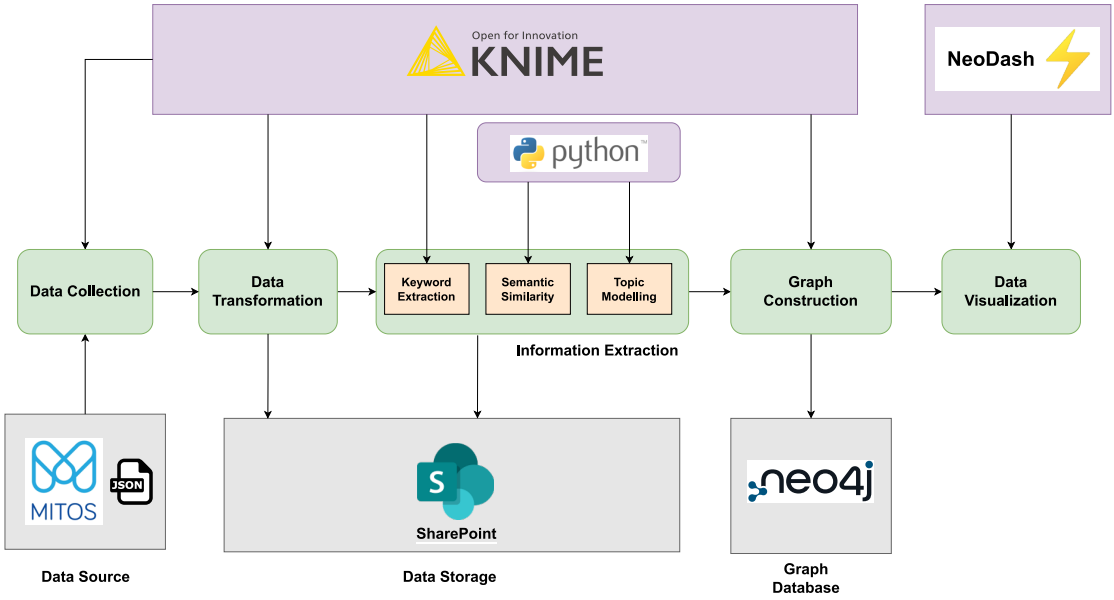
HAS_SEM_S□M□LAR_STEP	Σχέση μεταξύ βημάτων που δείχνει την ύπαρξη σημασιολογικής ομοιότητας μεταξύ των περιγραφών των βημάτων
HAS_SEM_S□M□LAR_STEP_D□G□TAL	Σχέση μεταξύ ψηφιακών βημάτων που δείχνει την ύπαρξη σημασιολογικής ομοιότητας μεταξύ των περιγραφών των ψηφιακών βημάτων

5 Τελική Αρχιτεκτονική και Εγκατάσταση

Στην παρούσα ενότητα περιγράφεται η τελική αρχιτεκτονική της υλοποίησης. Στη συνέχεια, περιγράφεται το εργαλείο NeoDash και οι οδηγίες εγκατάστασης των εργαλείων Neo4j και NeoDash.

5.1 Τελική Αρχιτεκτονική

Η τελική αρχιτεκτονική της υλοποίησης στα πλαίσια του πακέτου εργασίας WP5 παρουσιάζεται στην Εικόνα 11.



Εικόνα 11: Τελική αρχιτεκτονική WP5

Συνοπτικά, παρουσιάζονται τα βήματα της αρχιτεκτονικής:

- **Συλλογή δεδομένων (Data Collection)**: Γίνεται συλλογή δεδομένων από το MITOS API κάνοντας χρήση του εργαλείου KNIME.
- **Μετατροπή δεδομένων (Data Transformation)**: Χρησιμοποιείται το εργαλείο KNIME για να μετατρέψει τα δεδομένα του MITOS API από JSON σε μορφή πινάκων. Τα αποτελέσματα αποθηκεύονται σε αρχεία στο Microsoft SharePoint (ή εναλλακτικά οποιοδήποτε άλλο μέσο αποθήκευσης).
- **Εξαγωγή πληροφορίας (Information Extraction)**: Γίνεται επεξεργασία διάφορων πεδίων του ΕΜΔ με περιγραφές. Πιο συγκεκριμένα,

επιτυγχάνεται η εξαγωγή λέξεων-κλειδιών (Keyword Extraction) κάνοντας χρήση του εργαλείου KNIME και Python, καθώς και η εξαγωγή θεμάτων (Topic Modelling) και η εύρεση περιγραφών με σημασιολογική ομοιότητα (Semantic Similarity) κάνοντας χρήση Python. Τα αποτελέσματα της διεργασίας αποθηκεύονται στο Microsoft SharePoint.

- **Κατασκευή γράφου (Graph Construction)**: Γίνεται η κατασκευή του γράφου βασιζόμενο στο graph database Neo4j και χρησιμοποιείται το εργαλείο KNIME για την εισαγωγή των δεδομένων στο γράφο καθώς και την εφαρμογή queries για τον εμπλουτισμό του γράφου με περαιτέρω ανάλυση γράφου (graph analytics). Αξίζει να σημειωθεί πως το μέγεθος της βάσης δεδομένων του γράφου γνώσης που κατασκευάστηκε, ύστερα και από τον εμπλουτισμό με την επιπλέον πληροφορία που εξήχθησε και την πλούσια σύνδεση μεταξύ των κόμβων, ήταν πολλαπλάσια σε σχέση με το μέγεθος των δεδομένων που εξήχθησαν από το MITOS API. Πιο συγκεκριμένα το συνολικό μέγεθος της βάση δεδομένων του γράφου γνώσης ανήλθε περίπου στα 860MB, ενώ το μέγεθος των δεδομένων που αντλήθηκαν με όλη την πληροφορία του ΜΙΤΟΣ ήταν περίπου 65MB.
- **Οπτικοποίηση αποτελεσμάτων (Data Visualization)**: Χρησιμοποιείται το εργαλείο NeoDash για την επεξεργασία και δημιουργία dashboard που οπτικοποιεί τα αποτελέσματα σε μορφή γράφου, πινάκων και άλλων τρόπων οπτικοποίησης.

5.2 NeoDash

Το NeoDash[25] είναι ένα εργαλείο ανοιχτού κώδικα για τη δημιουργία οπτικοποιήσεων (dashboard) πάνω στο εργαλείο Neo4j. Παρέχει τη δυνατότητα δημιουργία διαδραστικών οπτικοποιήσεων σε μορφή γράφου αλλά και πινάκων και άλλων διαγραμμάτων. Ιδιαίτερα, η περίπτωση οπτικοποίησης των αποτελεσμάτων σε μορφή γράφου αποτελεί ένα πιο εύκολο και κατανοητό τρόπο για τους τελικούς χρήστες, όπου έχουν τη δυνατότητα να δουν το περιεχόμενο κάθε οντότητας (δηλαδή δημόσιων υπηρεσιών στα πλαίσια αυτού του

[25] https://github.com/neo4j-labs/neodash

παραδοτέου) μέσω των σχέσεων που παρουσιάζουν. Παράλληλα, το NeoDash παρουσιάζει σημαντικές δυνατότητες επαναχρησιμοποίησης και συνεργασίας, όπου το dashboard που επεξεργάζονται οι προγραμματιστές, μπορεί να εξαχθεί σε μορφή JSON και να διαμοιραστεί με άλλους. Το εργαλείο παρέχει έναν αριθμό από έτοιμα μέσα οπτικοποίησης (no-code), αλλά παρέχει και τη δυνατότητα υλοποίησης οπτικοποιήσεων με συγκεκριμένες προσαρμογές κάνοντας χρήση της γλώσσας Cypher (low-code) που ενσωματώνεται στο Neo4j και επιτρέπει την συλλογή πληροφοριών μέσα από μία βάση δεδομένων γράφου γνώσεως.

5.3 Εγκατάσταση και ρύθμιση εργαλείων και υποδομών

Στα πλαίσια πειραμάτων και υλοποίησης του συγκεκριμένου παραδοτέου, έγινε χρήση των υποδομών του ΕΔΥΤΕ μέσα από την πλατφόρμα Okeanos Knossos. Πιο συγκεκριμένα, χρησιμοποιήθηκε VM Ubuntu Linux Server 16.04 (16GB RAM, 16 CPUs, 130GB attached disk) το οποίο στη συνέχεια αναβαθμίστηκε στην έκδοση 20.04 για τη συμβατότητα με τα σχετικά εργαλεία.

Εγκατάσταση Neo4j

Το Neo4j Community Edition[26] αποτελεί εργαλείο ανοιχτού κώδικα και χρησιμοποιήθηκε για τη δημιουργία και ανάλυση του γράφου με δεδομένα του ΕΜΔΔ. Οι οδηγίες εγκατάστασης του εργαλείου (v. 5.11) σε Ubuntu Linux Server με τις απαιτούμενες επεκτάσεις και ρυθμίσεις περιγράφονται στο GitLab. Εναλλακτικός τρόπος εγκατάστασης αποτελεί η χρήση Docker Container[27].

Εγκατάσταση NeoDash

Ο ευκολότερος τρόπος εγκατάστασης του εργαλείου NeoDash είναι κάνοντας χρήση Docker. Το NeoDash χρησιμοποιείται σε δύο μορφές:

- Editor Mode: Χρησιμοποιείται από τους χρήστες που δημιουργούν και επεξεργάζονται το Dashboard, το οποίο μπορεί να εξαχθεί σε μορφή JSON και να επαναχρησιμοποιηθεί από άλλους.
- Standalone (Viewer) Mode: Χρησιμοποιείται από τους τελικούς χρήστες που καταναλώνουν το Dashboard, οι οποίοι δεν έχουν δυνατότητα

[26] https://neo4j.com/deployment-center/?ref=subscription#community
[27] https://neo4j.com/docs/operations-manual/current/docker/

επεξεργασίας αλλά μόνο προβολής. Για να επιτευχθεί αυτό, οι χρήστες που επεξεργάζονται το Dashboard, αποθηκεύουν το Dashboard μέσα στο Neo4j και μετά δημιουργείται καινούργιο Docker Container που επαναχρησιμοποιεί το αποθηκευμένο Dashboard.

Αναλυτικές οδηγίες εγκατάστασης του NeoDash στις παραπάνω μορφές παρέχονται στο GitLab.

6 Παρουσίαση Διαδραστικού Dashboard

Η ενότητα αυτή αφορά το τελευταίο στάδιο της μεθοδολογίας μας, στο οποίο πραγματοποιείται η παρουσίαση των αποτελεσμάτων της ανάλυσης του γράφου, μέσω ενός διαδραστικού dashboard. Τα αποτελέσματα παρουσιάζονται σε 5 διαφορετικές καρτέλες για την καλύτερη απεικόνιση και κατανόηση των αποτελεσμάτων, τα οποία αναλύονται στις παρακάτω υποενότητες.

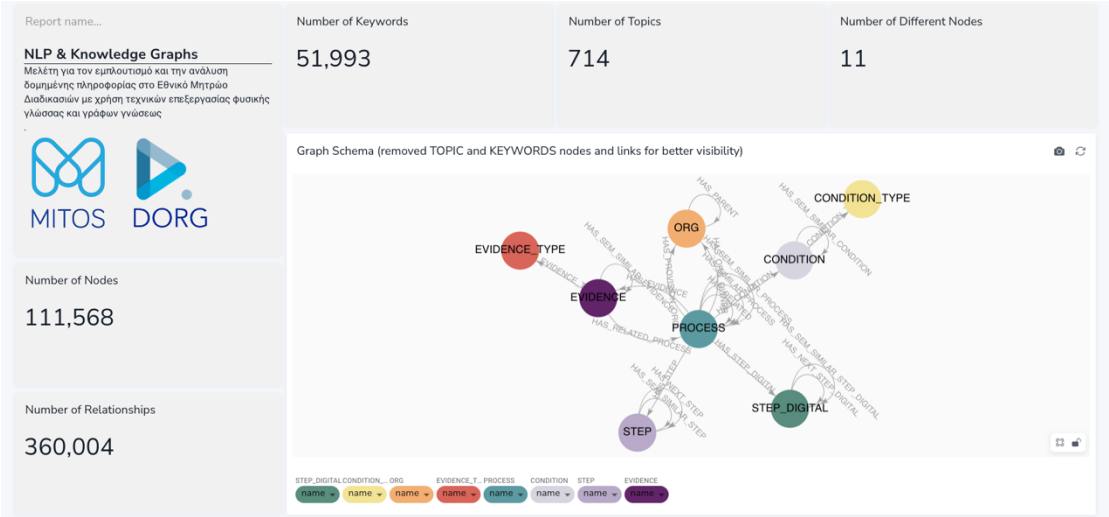
6.1 Graph Overview

Στην καρτέλα "Graph Overview", παρουσιάζεται μία γενική εικόνα του γράφου και της δομής αυτού, ώστε να υπάρχει καλύτερη κατανόηση από τους χρήστες των αναλύσεων που θα συναντήσουν στην συνέχεια.

Πιο συγκεκριμένα, και όπως αυτό απεικονίζεται στην Εικόνα 12, αρχικά παρουσιάζεται το βασικό σχήμα - δομή του γράφου. Βασικός κόμβος του γράφου είναι η διαδικασία *"PROCESS"*, ο οποίος συνδέεται με τα βήματα *"STEP"*, ψηφιακά βήματα *"STEP_DIGITAL"*, προϋποθέσεις "*CONDITION*", δικαιολογητικά "*EVIDENCE*" και οργανισμούς "*ORG*". Επίσης, τα δικαιολογητικά συνδέονται με τον τύπο του κάθε δικαιολογητικού "*EVIDENCE_TYPE*" και οι προϋποθέσεις με τον τύπο της προϋπόθεσης "*CONDITION_TYPE*". Παράλληλα, στους περισσότερους κόμβους υπάρχει και η εσωτερική σχέση της ομοιότητας μεταξύ των κόμβων, στα πλαίσια της ομογενοποίησης που παρουσιάστηκε στην ενότητα 3.3 και αποτέλεσε τον κύριο παράγοντα συνδεσιμότητας του γράφου. Για παράδειγμα δύο βήματα όπου το πρώτο είναι *«Εκκίνηση Διαδικασίας»* και το δεύτερο *«Εκκίνηση της Διαδικασίας»* θα έχουν την σχέση *"HAS_SEM_SIMILAR_STEP"* , καθώς είναι σημασιολογικά κοντά. Επίσης, όσον αφορά τα βήματα έχει καταγραφή και η σχέση *"HAS_NEXT_STEP"*, το οποίο υποδηλώνει το επόμενο βήμα μέσα σε μία διαδικασία.

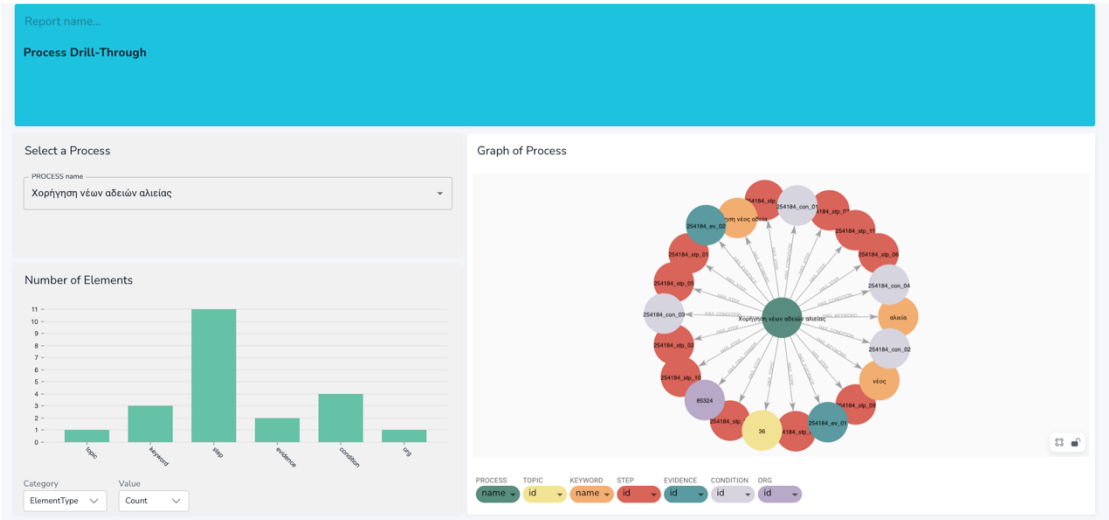
Τέλος, αξίζει να σημειωθεί, πως οι κόμβοι *"KEYWORD"* και *"TOPIC"* δεν εμφανίζονται στο διάγραμμα, καθώς η οπτικοποίηση του γράφου θα ήταν αρκετά πιο πολύπλοκη και θα είχε αντίθετα αποτελέσματα από τον σκοπό που είναι η διευκόλυνση του χρήστη. Παρόλα αυτά, εμφανίζονται οι πληροφορίες των συγκεκριμένων κόμβων στα στατιστικά που παρουσιάζονται στην εικόνα.

Επίσης, δύο ενδιαφέροντα στοιχεία από τα στατιστικά που απεικονίζονται, είναι ο αριθμός των κόμβων που υπάρχουν στον γράφο (111.568) και οι σχέσεις που υπάρχουν μεταξύ αυτών των κόμβων (360.004), γεγονός που υποδηλώνει ένα αρκετά σύνθετο γράφο γνώσης, με πολλές συνδέσεις μεταξύ των στοιχείων του που χρήζει περαιτέρω ανάλυσης.



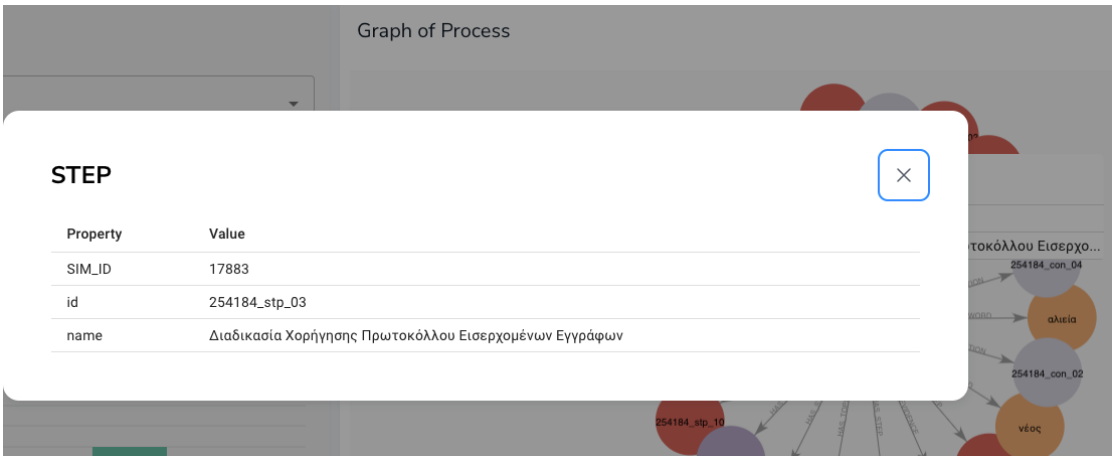
Εικόνα 12: Σχήμα του Γράφου και στατιστικά των επιπλέον Κόμβων

Αφού ο χρήστης έχει πλέον καλύτερη κατανόηση της δομής του γράφου γνώσης, στη συνέχεια αυτής της υποενότητας, έχει την δυνατότητα να επιλέξει μία διαδικασία και να αλληλοεπιδράσει μαζί της. Πιο συγκεκριμένα, όπως παρουσιάζεται στην Εικόνα 13, επιλέγοντας οποιαδήποτε διαδικασία, θα εμφανιστεί ένας γράφος, που παρουσιάζει αυτές τις συνδέσεις με τον βασικό κόμβο, δηλαδή τη διαδικασία *"PROCESS"*, παρέχοντας έτσι μια καλή εικόνα για το περιεχόμενο της επιλεγμένης διαδικασίας. Παράλληλα, θα εμφανιστούν στον χρήστη μέσω ενός bar chart μερικά στατιστικά σχετικά με τον αριθμό των στοιχείων του κάθε κόμβου που συνδέονται με αυτή, όπως για παράδειγμα ο αριθμός των βημάτων.



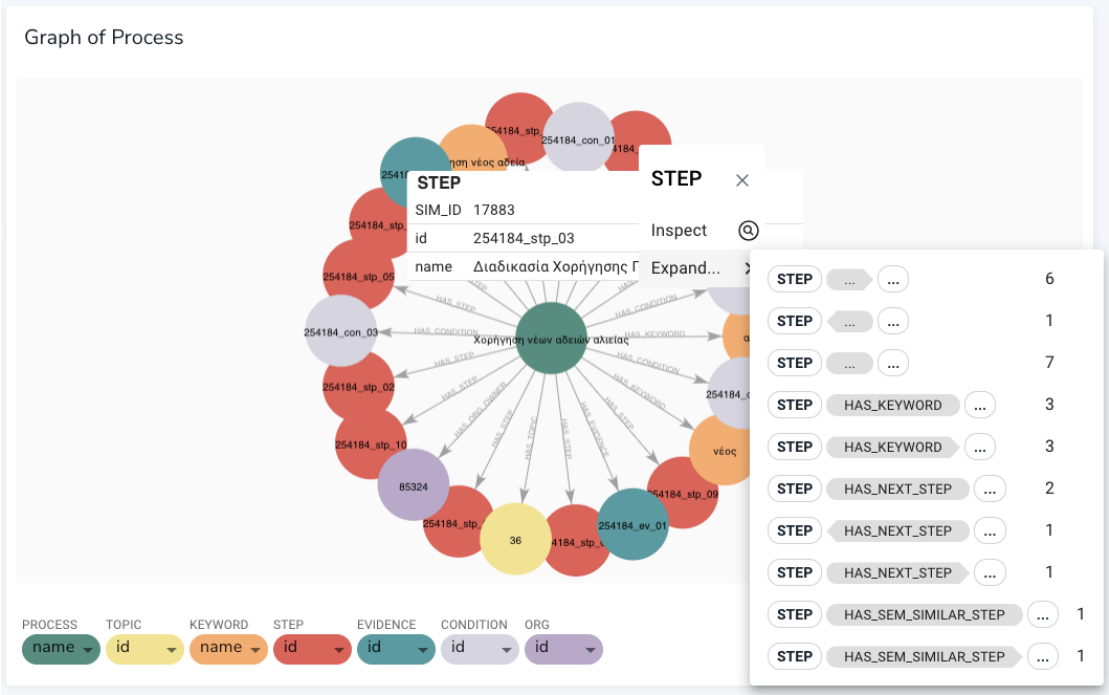
Εικόνα 13: Παρουσίαση Στατιστικών και Γράφου μίας Διαδικασίας

Σε αυτό το σημείο, αξίζει να αναφερθούν κάποιες λειτουργίες του dashboard, όπου μπορούν να εφαρμοστούν σε κάθε γράφο γνώσης που απεικονίζεται στην συνέχεια της ανάλυσής μας. Αρχικά, κάνοντας αριστερό κλικ σε έναν κόμβο του γράφου, στην συγκεκριμένη περίπτωση ενός βήματος, εμφανίζεται ένα αναδυόμενο παράθυρο το οποίο δίνει περισσότερα στοιχεία για αυτόν τον συγκεκριμένο κόμβο (Εικόνα 14).



Εικόνα 14: Αναδυόμενο παράθυρο με πληροφορίες του Κόμβου

Στην περίπτωση που ο χρήστης θέλει να εξερευνήσει περαιτέρω σχέσεις που υπάρχουν για ένα κόμβο, αλλά δεν εμφανίζονται στο γράφο, τότε, κάνοντας δεξί κλικ και επιλέγοντας *"Expand"* έχει την δυνατότητα να εμφανίσει όλες της διαθέσιμες επιλογές για αυτόν τον κόμβο (Εικόνα 15).

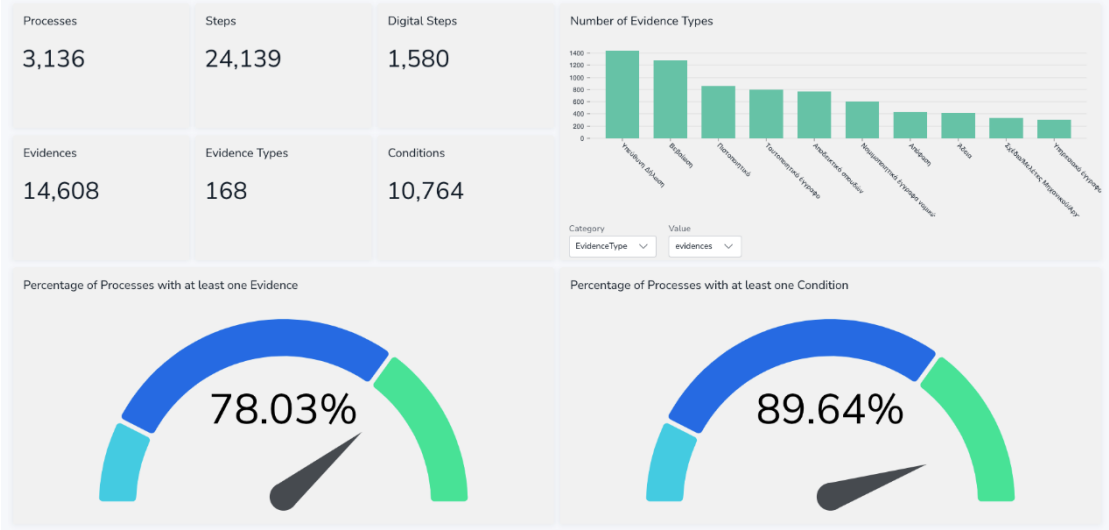


Εικόνα 15: Εμφάνιση επιπλέον σχέσεων ενός κόμβου

6.2 General Statistics

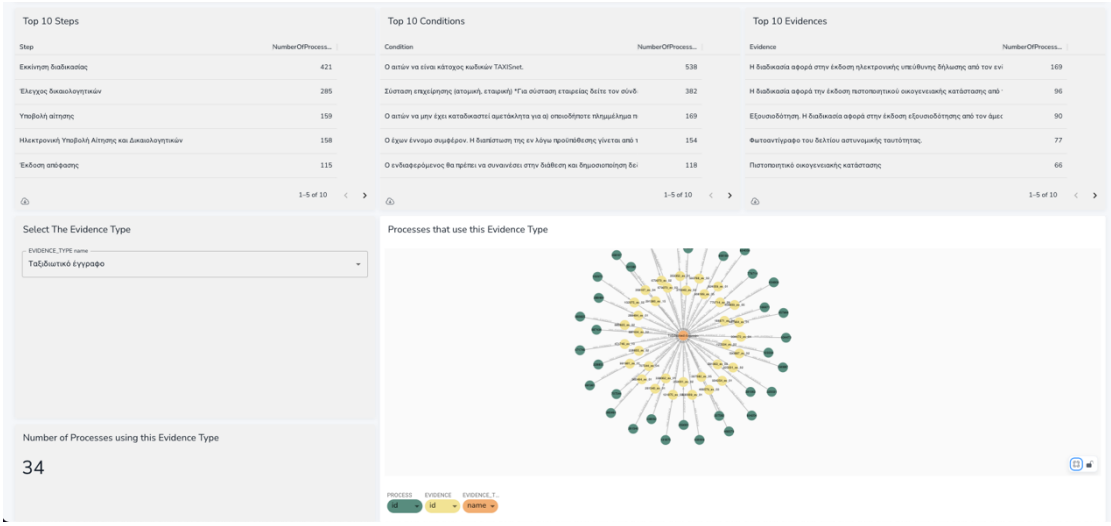
Συνεχίζοντας την ανάλυσή μας προχωράμε στην καρτέλα “General Statistics”, όπου παρουσιάζονται μερικά γενικά στατιστικά γύρω από τα δεδομένα του ΜΙΤΟΣ. Τα συγκεκριμένα στατιστικά βασίζονται κυρίως σε πληροφορία που επιστρέφεται από το ΜΙΤΟΣ API.

Αναλυτικότερα, αρχικά στην καρτέλα “General Statistics” (Εικόνα 16) παρουσιάζονται μερικά γενικά στοιχεία των δεδομένων μας, όπως για παράδειγμα ο αριθμός των διαδικασιών που είναι 3.136. Στην συνέχεια, υπάρχει ένα bar chart, το οποίο απεικονίζει τον αριθμό των 10 πιο συχνών τύπων δικαιολογητικών που συναντώνται στις διαδικασίες, με την υπεύθυνη δήλωση να είναι ο πιο συχνός τύπος δικαιολογητικού, καθώς ζητείται 1439 φορές. Τέλος μέσω δύο gauge charts παρουσιάζεται το ποσοστό των διαδικασιών που ζητούν τουλάχιστον ένα δικαιολογητικών (78,03%) και έχουν τουλάχιστον μία προϋπόθεση για να εκτελεστούν (89,64%).



Εικόνα 16: Γενικά στατιστικά σχετικά με τα δεδομένα από το ΜΙΤΟΣ API

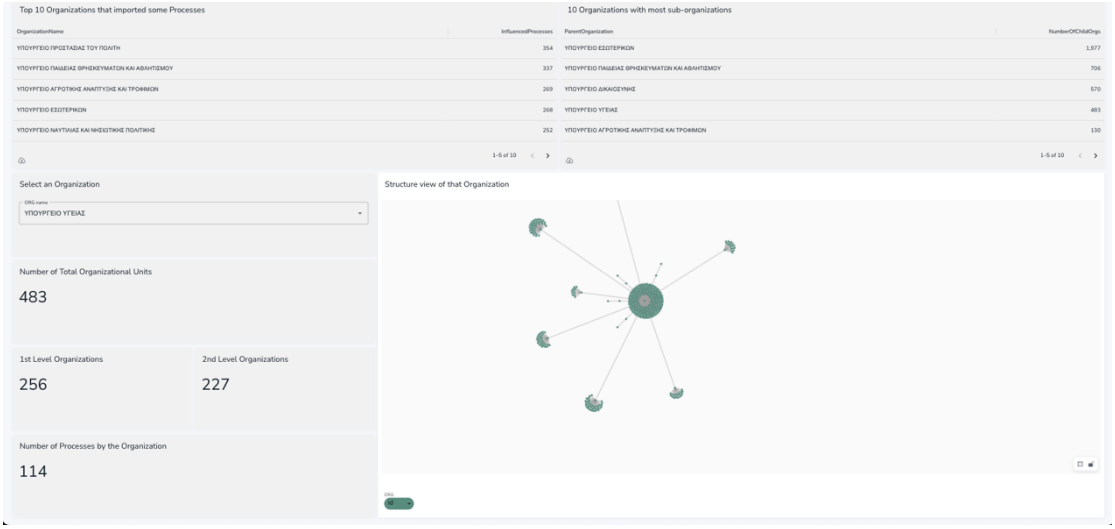
Ακολούθως, παρουσιάζονται μερικά γενικά στοιχεία σχετικά με τα Δικαιολογητικά, Προϋποθέσεις και τα Βήματα που εξάγονται εφόσον έχει πραγματοποιηθεί ομογενοποίηση των δεδομένων (Εικόνα 17). Πιο συγκεκριμένα, παρουσιάζονται σε πίνακες, οι οποίοι μπορούν επίσης να κατέβουν ως CSV, τα 10 Δικαιολογητικά, 10 Προϋποθέσεις και 10 Βήματα που χρησιμοποιούνται πιο πολλές φορές από τις διαδικασίες. Στην συνέχεια ο χρήστης έχει την δυνατότητα να επιλέξει ένα τύπο δικαιολογητικού "*EVIDENCE_TYPE*" και να δει σε έναν διαδραστικό γράφο, τα δικαιολογητικά και τις διαδικασίες που συνδέονται με αυτό τον τύπο δικαιολογητικού, καθώς επίσης και τον αριθμό των διαδικασιών που χρησιμοποιούν αυτό τον τύπο δικαιολογητικού.



Εικόνα 17: Γενικά στοιχεία σχετικά με τα Δικαιολογητικά, Προϋποθέσεις και Βήματα

Στο τέλος της καρτέλας, παρουσιάζονται μερικά γενικά στοιχεία όσον αφορά τους οργανισμούς (Εικόνα 18). Πιο συγκεκριμένα, στην αρχή υπάρχουν δύο πίνακες, όπου στον έναν εμφανίζονται οι 10 οργανισμοί που έχουν καταχωρίσει τις περισσότερες διαδικασίες στο ΜΙΤΟΣ, ενώ στον άλλον οι 10 οργανισμοί με τους περισσότερους υπο-οργανισμούς με το Υπουργείο Εσωτερικών να είναι πρώτο με 1977 υπο-οργανισμούς και με μεγάλη διαφορά από τον επόμενο οργανισμό.

Τέλος, ο χρήστης έχει την δυνατότητα να επιλέξει έναν οργανισμό και να δει μερικά στατιστικά στοιχεία σχετικά με αυτόν, όπως για παράδειγμα τον αριθμό των διαδικασιών που είναι καταχωρημένες στο ΜΙΤΟΣ. Επίσης, εμφανίζεται ένας γράφος που παρουσιάζει τις σχέσεις με τους υπο-οργανισμούς (διάρθωση). Αυτή η απεικόνιση, μπορεί να βοηθήσει τον χρήστη να κατανοήσει την πολυπλοκότητα του εγχειρήματος του ΜΙΤΟΣ σχετικά με την καταγραφή των διαδικασιών του δημοσίου, καθώς οπτικοποιείται το βάθος των δομών και των συσχετίσεων των δημόσιων οργανισμών.



Εικόνα 18: Γενικά στοιχεία σχετικά με τους Οργανισμούς

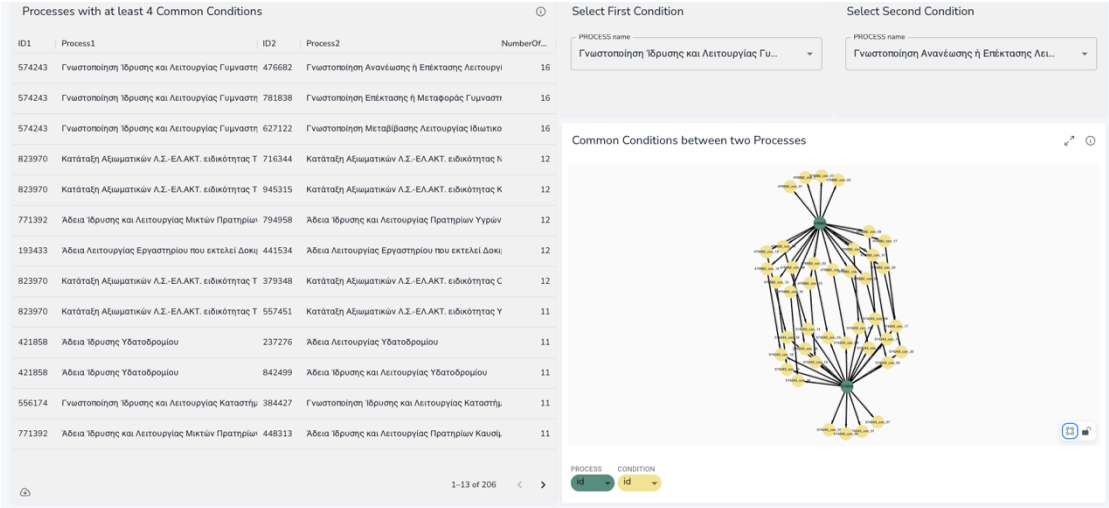
6.3 Graph Analysis through common Nodes

Στην συνέχεια, οι αναλύσεις που παρουσιάζονται, βασίζονται σε δεδομένα που εξήχθησαν με τεχνικές επεξεργασίας φυσικής γλώσσας, μηχανικής μάθησης και ανάλυσης γράφου. Ειδικότερα, στην καρτέλα “Graph Analysis through common Nodes”, παρουσιάζονται τα αποτελέσματα εντοπισμού κοινών στοιχείων μεταξύ των διαδικασιών. Τα στοιχεία που εξετάζονται είναι οι προϋποθέσεις “*CONDITION*”, τα δικαιολογητικά “*EVIDENCE*”, συνδυασμός αυτών καθώς και τα βήματα “STEP”. Τα αποτελέσματα των παραπάνω αναλύσεων παρουσιάζονται με παρόμοιο τρόπο στο dashboard, επομένως στο παραδοτέο θα επεξηγηθεί μόνο η ανάλυση για εύρεση κοινών προϋποθέσεων μεταξύ των διαδικασιών.

Όπως παρουσιάζεται στην Εικόνα 19, υπάρχει ένας πίνακας, όπου καταγράφονται διαδικασίες με κοινές περιγραφές προϋποθέσεων. Πιο συγκεκριμένα, σε κάθε γραμμή υπάρχουν τα ID και τα ονόματα των δύο διαδικασιών και στο τελευταίο κελί εμφανίζεται ο αριθμός των κοινών προϋποθέσεων μεταξύ αυτών των διαδικασιών. Στη συνέχεια, ο χρήστης έχει την δυνατότητα μέσω δύο drop-down μενού, να επιλέξει τα ζεύγη των διαδικασιών που θέλει να δει αναλυτικότερα, και να εμφανιστεί το ανάλογο διάγραμμα, όπου εμφανίζονται οι δύο διαδικασίες με τις προϋποθέσεις τους και οι προϋποθέσεις που συνδέονται μεταξύ τους, εμφανίζουν τις ανάλογες συνδέσεις. Όπως

αναφέρθηκε και προηγουμένως, παρόμοια λογική ακολουθείται και στις υπόλοιπες περιπτώσεις που εξετάζονται.

Η ανάδειξη τέτοιων ζευγαριών διαδικασιών, μπορεί να αποτελέσει ένα σημαντικό εργαλείο στα χέρια των ειδικών για εξέταση τυχόν λαθών στις καταχωρήσεις κάποιων διαδικασιών. Για παράδειγμα, μία διαδικασία που έχει όλα τα βήματα της κοινά με μία άλλη διαδικασία, αλλά η δεύτερη διαδικασία έχει επιπλέον ένα βήμα, μήπως είναι περιττό ή μήπως αυτό το βήμα θα έπρεπε να ενσωματωθεί και στην πρώτη διαδικασία;



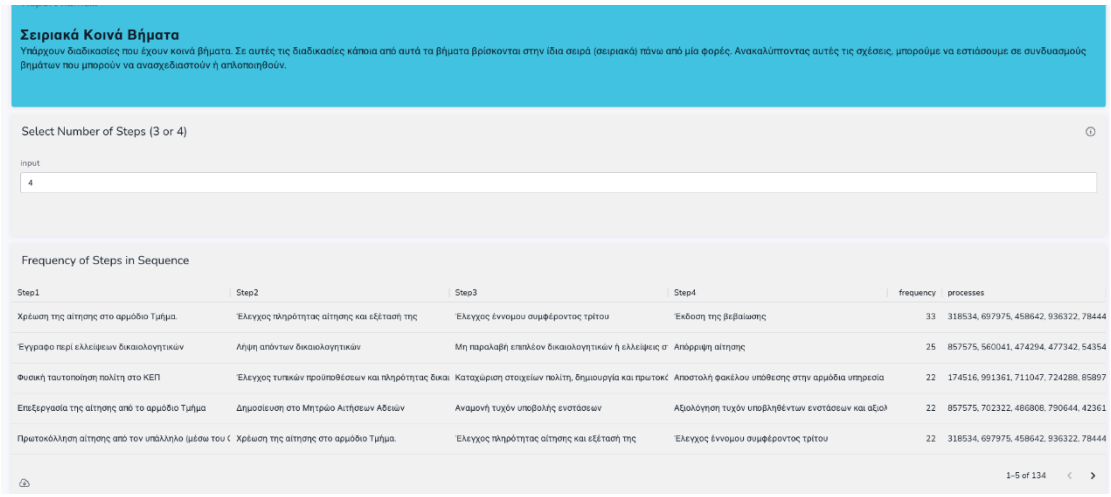
Εικόνα 19: Διαδικασίες με κοινές περιγραφές Προϋποθέσεων

6.4 Advanced Graph Analysis

Στην επόμενη καρτέλα “Advanced Graph Analysis”, όπως παραπέμπει και ο τίτλος της, παρουσιάζονται πιο σύνθετες αναλύσεις με τα δεδομένα του γράφου. Αρχικά, παρουσιάζονται τα σειριακά κοινά βήματα εντός των διαδικασιών, δηλαδή βήματα που βρίσκονται στην ίδια σειρά (σειριακά) πάνω από μία φορές. Αναλυτικότερα, όπως παρουσιάζεται και στην Εικόνα 20 ο χρήστης έχει την δυνατότητα επιλογής για εμφάνιση τριών ή τεσσάρων σειριακών βημάτων. Με βάση την επιλογή αυτή, δημιουργείται ένας πίνακας, όπου καταγράφονται τα σειριακά βήματα με ένα αντιπροσωπευτικό όνομα, η συχνότητα με την οποία συναντιούνται, δηλαδή ο συνολικός αριθμός των διαδικασιών που έχουν

χρησιμοποιούν αυτή την σειρά βημάτων, καθώς και τα ID των διαδικασιών αυτών. Φυσικά, υπάρχει η δυνατότητα λήψης αυτού του πίνακα σε μορφή CSV.

Αξίζει να σημειωθεί, πως τα αποτελέσματα αυτής της ανάλυσης, μπορούν να αποτελέσουν εφαλτήριο ώστε να ανακαλυφθούν συνδυασμοί βημάτων που μπορούν να ανασχεδιαστούν ή να απλοποιηθούν. Επίσης, μπορεί να αποτελέσει μία βάση για την δημιουργία ενός recommendation engine βημάτων.

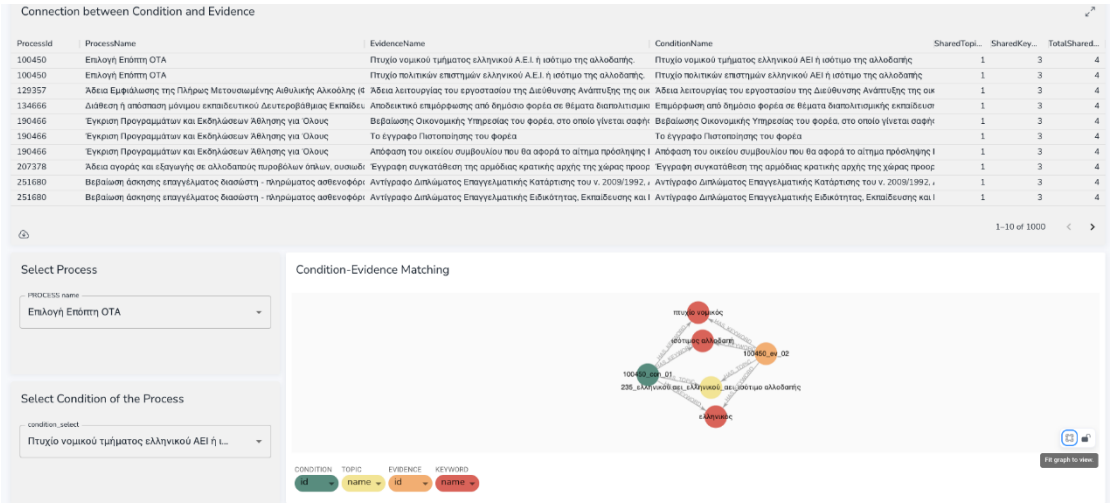


Εικόνα 20: Σειριακά Κοινά Βήματα

Στην ίδια καρτέλα εξετάζεται η ύπαρξη σχέσεων μεταξύ δικαιολογητικών και προϋποθέσεων μίας διαδικασίας. Για να επιτευχθεί αυτό, ελήφθησαν υπόψιν οι κόμβοι Topics και Keywords ως συνδετικοί κρίκοι των δικαιολογητικών και προϋποθέσεων. Ειδικότερα, στο dashboard υπάρχει ένας πίνακας ο οποίος παρουσιάζει όλες τις 1000 σχέσεις που εξήχθησαν από την ανάλυση (Εικόνα 21). Στις πρώτες δύο στήλες εμφανίζονται το ID και το όνομα μία διαδικασίας και οι δύο επόμενες στήλες παρουσιάζουν ένα δικαιολογητικό και μία προϋπόθεση αυτής. Οι τρεις τελευταίες στήλες καταγράφουν τον αριθμό των topics, των keywords και του αθροίσματος αυτών που είναι κοινά μεταξύ του ζεύγους που εξετάζεται. Αν ο χρήστης θέλει να δει και να εξετάσει περεταίρω ένα ζεύγος βλέποντας ποια keywords και topics τα συνδέουν, έχει την δυνατότητα να το

πραγματοποιήσει μέσω ενός διαδραστικού γράφου, επιλέγοντας αρχικά την διαδικασία και στην συνέχεια την προϋπόθεση που επιθυμεί να εξεταστεί.

Η ανακάλυψη της σύνδεσης μεταξύ προϋποθέσεων και δικαιολογητικών μπορεί να οδηγήσει στην προτεραιοποίηση διαδικασιών που μπορούν να εξεταστούν για την μεταφορά τους από “Document Driven” σε “Data Driven” διαδικασίες, καθώς υπάρχει μία βάση ώστε να εντοπιστεί το στοιχείο που πρέπει να εξαχθεί μέσα από το δικαιολογητικό που απαιτείται για να ικανοποιήσει μία συγκεκριμένη συνθήκη.



Εικόνα 21: Σχέσεις μεταξύ Προϋποθέσεων και Δικαιολογητικών

6.5 Process Similarity and Clustering

Στην τελευταία καρτέλα “Process Similarity and Clustering”, παρουσιάζονται αποτελέσματα που εξήχθησαν με την χρήση αλγορίθμων γράφων. Όπως παρουσιάστηκε τόσο στο προηγούμενο παραδοτέο όσο και στην παρουσίαση του dashboard στις παραπάνω καρτέλες, η δομή του γράφου και οι πλούσιες συνδέσεις μεταξύ των κόμβων δίνει την δυνατότητα εξαγωγής αρκετών χρήσιμων συμπερασμάτων χρησιμοποιώντας κάποια queries. Οι δυνατότητες δεν περιορίζονται όμως μόνο σε αυτό το κομμάτι. Μέσω της χρήσης αλγορίθμων γράφων, μπορούν να εξαχθούν και πιο σύνθετα αποτελέσματα.

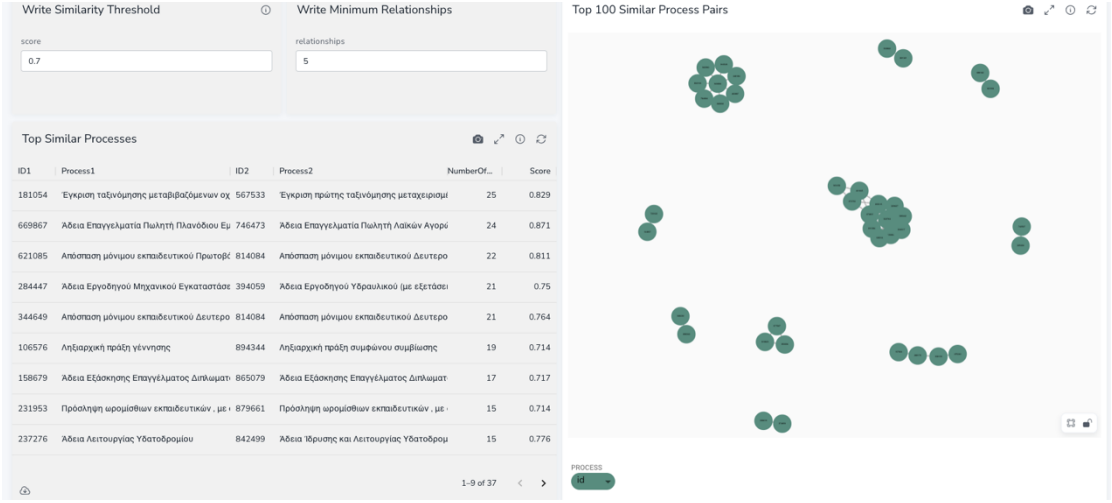
Η ανάλυση που παρουσιάζεται αρχικώς στο dashboard, αφορά την ομοιότητα μεταξύ των διαδικασιών μέσω του αλγορίθμου Jaccard Similarity. Ειδικότερα, η πληροφορία που λαμβάνεται υπόψιν είναι:

i. Οι διαδικασίες καθώς και τα Cluster IDs που θα παρουσιαστούν ακολούθως
ii. Οι σχέσεις με τους νέους κόμβους Βημάτων και τα Ψηφιακών Βημάτων
iii. Οι σχέσεις με τους νέους κόμβους Δικαιολογητικών
iv. Οι σχέσεις με τους νέους κόμβους Προϋποθέσεων
v. Οι σχέσεις με τους κόμβους Topics

Εφόσον πραγματοποιηθεί η ανάλυση, το αποτέλεσμα της ομοιότητας μεταξύ των διαδικασιών καταγράφεται σαν πληροφορία στον γράφο και στην συνέχεια υπάρχει η δυνατότητα παρουσίασης των αποτελεσμάτων, όπως αυτά εμφανίζονται στην Εικόνα 22. Πιο συγκεκριμένα, ο χρήστης έχει την δυνατότητα επιλέγοντας το ελάχιστο ποσοστό ομοιότητας μεταξύ των διαδικασιών, καθώς και τον αριθμό των ελάχιστων σχέσεων που αυτές οι διαδικασίες πρέπει να έχουν μεταξύ τους, να εμφανίσει αυτές που πληρούν αυτά τα κριτήρια.

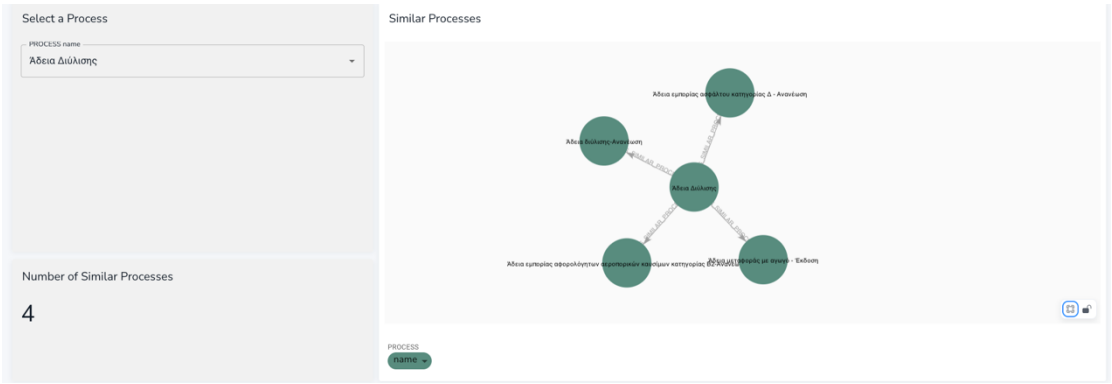
Παράλληλα, παρουσιάζονται μέσω ενός διαδραστικού γράφου τα 100 πιο όμοια ζεύγη διαδικασιών. Όπως είναι εμφανές, υπάρχουν κάποιες ομάδες διαδικασιών. Το ιδιαιτέρως ενδιαφέρον αυτών των ομάδων είναι πως αποτελούνται από διαδικασίες που ανήκουν στον ίδιο οργανισμό, γεγονός που υποδηλώνει την πιθανότητα ύπαρξης bias στον τρόπο που καταγράφονται οι διαδικασίες εντός του κάθε φορέα. Αυτό αποτελεί ένα ενδιαφέρον στοιχείο για περαιτέρω αξιολόγηση.

Επιπλέον, η πληροφορία της ομοιότητας μεταξύ των διαδικασιών μπορεί να ενσωματωθεί στο ΜΙΤΟΣ, ώστε να υπάρχουν περισσότερες συνδέσεις μεταξύ των διαδικασιών, με σκοπό την πιο εύκολη περιήγηση του χρήστη.



Εικόνα 22: Ομοιότητα μεταξύ των Διαδικασιών

Επιπρόσθετα, στο ίδιο πλαίσιο, εξετάζεται η δυνατότητα απεικόνισης παρόμοιων διαδικασιών με βάση μία διαδικασίας που θα επιλεχθεί από ένα drop-down μενού (Εικόνα 23).

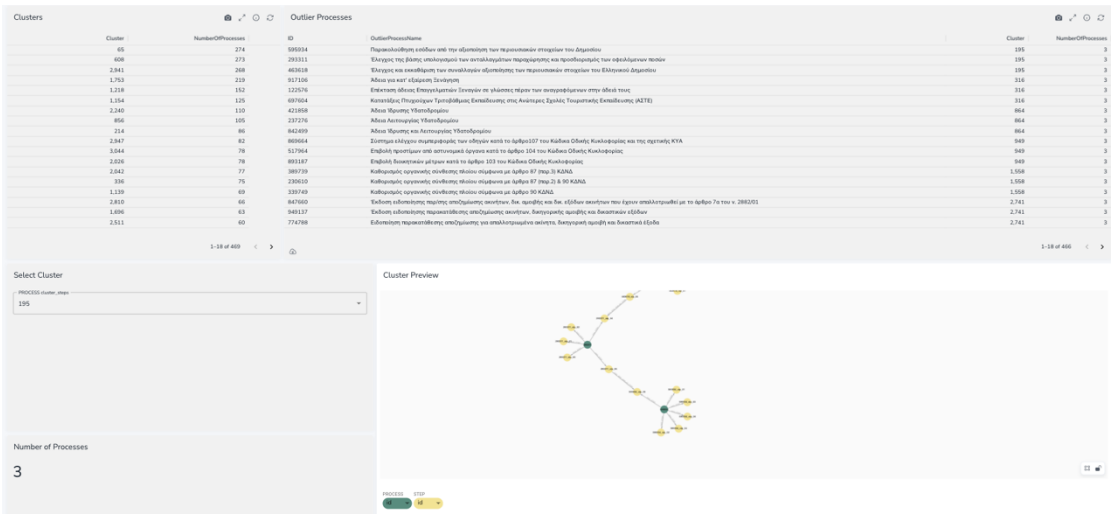


Εικόνα 23: Παρόμοιες Διαδικασίες μία Διαδικασίας

Η τελευταία ανάλυση που παρουσιάζεται στο dashboard αφορά τις ομαδοποιήσεις (clusters) των διαδικασιών με βάση τα βήματα. Όπως είχε παρουσιαστεί στο προηγούμενο παραδοτέο, εφαρμόζοντας τον αλγόριθμο Louvain στα βήματα, και αξιοποιώντας πάντα την πληροφορία της ομογενοποίησης, εντοπίζονται κοινότητες διαδικασιών. Αυτή η πληροφορία διατηρείται στον γράφο και υπάρχει η δυνατότητα παρουσίασης των αποτελεσμάτων που εμφανίζονται στην Εικόνα 24. Ειδικότερα, Υπάρχουν δύο

πίνακες, όπου στον ένα αποτυπώνεται ο αριθμός των διαδικασιών που ανήκουν σε κάθε cluster. Στον δεύτερο πίνακα καταγράφονται οι διαδικασίες οι οποίες ανήκουν σε cluster μικρότερο των 4 διαδικασιών, όντας πιθανών outliers. Φυσικά, όπως και σε προηγούμενες περιπτώσεις, υπάρχει η δυνατότητα μέσω επιλογής ενός cluster, να απεικονιστεί σε έναν γράφο.

Η ομαδοποίηση των διαδικασιών, μπορεί να αποτελέσει χρήσιμη πληροφορία για την προτεραιοποίηση των διαδικασιών για ανασχεδιασμό, καθώς μία αλλαγή μπορεί να επιφέρει αλλαγές σε πολλαπλές διαδικασίες. Αντίστοιχα, οι διαδικασίες που χαρακτηρίζονται ως outlier μπορούν να χαρακτηριστούν ως αυτές που δεν χρήζουν προτεραιοποίησης για ανασχεδιασμό.



Εικόνα 24: Ομαδοποιήσεις (Clusters) Διαδικασιών

7 Συμπεράσματα & Κατευθύνσεις προς βελτίωση και περαιτέρω ανάπτυξη του ΕΜΔΔ

7.1 Συμπεράσματα

Στο συγκεκριμένο έργο το οποίο είχε ως στόχο την ενίσχυση της πληροφορίας στο ΕΜΔΔ μέσω της δημιουργίας ενός γράφου γνώσης με πλούσιες συνδέσεις μεταξύ των δεδομένων που παρέχονται από το ίδιο το ΕΜΔΔ πραγματοποιήθηκε αρχικά (1) άντληση των δεδομένων από το ΜΙΤΟΣ ΑΡ και (2) εξαγωγή επιπλέον δομημένων σχέσεων από αυτά μέσω σύγχρονων τεχνολογιών βαθιάς μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας. Παράλληλα, (3) έγινε πλήρης κατανόηση των δεδομένων του ΕΜΔΔ με σκοπό την (4) δημιουργία ροής εργασιών για τον διαχωρισμό της πληροφορίας που χρειάζεται για την κατασκευή του γράφου γνώσης. 'Ολες οι παραπάνω ενέργειες είχαν ως τελικό στόχο (5) την ανάλυση των δεδομένων μέσω της υψηλής διασύνδεσης της πληροφορίας που προσφέρει ο γράφος γνώσης για την εξαγωγή χρήσιμων αποτελεσμάτων που είτε είναι δύσκολο να αποκτηθεί με την χρήση σχεσιακών βάσεων, είτε ακόμη και να εντοπιστεί. Τα αποτελέσματα της επεξεργασίας και ανάλυσης των δεδομένων μπορούν να οδηγήσουν στα εξής χρήσιμα συμπεράσματα:

- Η χρήση σύγχρονων τεχνολογιών τεχνητής νοημοσύνης και επεξεργασίας φυσικής γλώσσας μπορεί να συντελέσει σημαντικά στην βελτίωση της ποιότητας των δεδομένων και την εξαγωγή πλούσιας δομημένης πληροφορίας όπως η ομογενοποίηση των βημάτων/δικαιολογητικών/προϋποθέσεων.
- Η ομογενοποίηση των δεδομένων (semantic similarity), όπως για παράδειγμα βημάτων που είναι γραμμένα με διαφορετικό τρόπο και η διατήρησης αυτής της σύνδεσης μέσω σχέσεων μέσα στον γράφο αποτελεί πολύτιμη πληροφορία για την εξαγωγή αποτελεσμάτων. Τα αποτελέσματα μπορεί να είναι τόσο βασικά στατιστικά, όπως τα 5 βήματα που χρησιμοποιούνται περισσότερο από τις διαδικασίες, όσο και η εύρεση ομάδων διαδικασιών με κοινά βήματα. Η πρότυπη λύση που έχει

εφαρμοστεί στα πλαίσια του παραδοτέου θα μπορούσε να αποτελέσει ένα εργαλείο υποβοήθησης των ειδικών του ΕΜΔΔ για την πλήρη ομογενοποίηση των δεδομένων.

- Η χρήση τεχνολογιών γράφου γνώσεως δίνει τη δυνατότητα εφαρμογής σύνθετων αναλύσεων μέσω των σχέσεων των δεδομένων, καθώς και παρέχει δυνατότητες οπτικοποίησης σε μορφή γράφου που αποτελούν έναν πιο κατανοητό τρόπο μετάδοσης της γνώσης προς τους τελικούς χρήστες.
- Επίσης, οι αναλύσεις που παρουσιάστηκαν τόσο στα πλαίσια του προηγούμενου παραδοτέου όσο και αυτού κατά την παρουσίαση του dashboard, αναδεικνύει στοιχεία που θα μπορούσαν να αποτελέσουν εναρκτήρια βάση για την προτεραιοποίηση:
  - Ανασχεδιασμού μίας διαδικασίας.
  - Μεμονωμένων ή συνδυασμό βημάτων που θα μπορούσαν να αυτοματοποιηθούν και θα είχαν αντίκτυπο σε μεγάλο εύρος διαδικασιών.
  - Δικαιολογητικών που με βάση την χρήση τους από διαδικασίες θα υπήρχε όφελος από την ψηφιοποίηση τους.
  - Διαδικασιών που θα μπορούσαν να μετατραπούν από “document centric” σε “data centric” καθώς έχουμε κάποιες αρχικές συνδέσεις μεταξύ δικαιολογητικών και προϋποθέσεων.
- Αντίστοιχα οι αναλύσεις που εξάγουν χρήσιμα συμπεράσματα αποτελούν πληροφορία που θα μπορούσε να αξιοποιηθεί ή να αποτελέσει αντικείμενο περαιτέρω διερεύνησης όπως:
  - Καταγραφή συνδέσεων μεταξύ διαδικασιών που αυτή την στιγμή δεν υπάρχουν.
  - Πιθανή ανάδειξη νέων κατηγοριοποιήσεων για τις διαδικασίες για πιο εύκολη πλοήγηση του χρήστη.

- Πιθανή ανάδειξη biased στοιχείων, όπως η πολύ μεγάλη ομοιότητα διαδικασιών μεταξύ του ίδιου φορέα.
- Πιθανή ανάδειξη στοιχείων ομοιότητας στις διαδικασίες μεταξύ δύο ή περισσοτέρων φορέων.

- Τέλος, αξίζει να σημειωθεί πως οι αναλύσεις που παρουσιάζονται στο διαδραστικό dashboard καθώς και η χρηστικότητα αυτού, μπορεί να αποτελέσει ένα σημαντικό εργαλείο στα χέρια των ειδικών του ΕΜΔΔ και να ανακαλυφθούν περιπτώσεις χρήσης οι οποίες δεν έχουν αναφερθεί στα πλαίσια των παραδοτέων του έργου.

Συνοψίζοντας, η εγγενής μορφή συσχετίσεων των κόμβων στους γράφους γνώσης και ο εμπλουτισμός της πληροφορίας με επιπλέον δομημένη πληροφορία που εξήχθησε, εμφανίζουν μία επιπλέον δυναμική στην εξόρυξη πληροφορίας από τα στοιχεία που υπάρχουν στο ΜΙΤΟΣ. Στοιχεία που με επιπλέον επεξεργασία μπορούν να οδηγήσουν σε πρόσθετες χρήσιμες σχέσεις που δεν είναι εμφανείς εκ πρώτης όψεως.