



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

**ΑΝΑΠΤΥΞΗ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΣΤΑΤΙΣΤΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ
ΜΕΤΑ-ΔΕΔΟΜΕΝΩΝ ΕΠΙΣΤΗΜΟΝΙΚΩΝ ΔΗΜΟΣΙΕΥΣΕΩΝ**

Διπλωματική Εργασία

Μυλωνάς Γεώργιος

Επιβλέπουσα: Δασκαλοπούλου Ασπασία

Ιούλιος 2023



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

**ΑΝΑΠΤΥΞΗ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΣΤΑΤΙΣΤΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ
ΜΕΤΑ-ΔΕΔΟΜΕΝΩΝ ΕΠΙΣΤΗΜΟΝΙΚΩΝ ΔΗΜΟΣΙΕΥΣΕΩΝ**

Διπλωματική Εργασία

Μυλωνάς Γεώργιος

Επιβλέπουσα: Δασκαλοπούλου Ασπασία

Ιούλιος 2023



UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

**DEVELOPMENT OF DATABASE FOR THE STATISTICAL ANALYSIS OF
SCIENTIFIC PUBLICATIONS METADATA**

Diploma Thesis

Mylonas Georgios

Supervisor: Daskalopoulou Aspasia

July 2023

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων/πουσα **Δασκαλοπούλου Ασπασία**

Αναπληρώτρια Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και
Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Σταμπουλής Γεώργιος**

Αναπληρωτής Καθηγητής, Τμήμα Οικονομικών Επιστημών,
Πανεπιστήμιο Θεσσαλίας

Μέλος **Βασιλακόπουλος Μιχαήλ**

Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών
Υπολογιστών

**ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ
ΔΙΚΑΙΩΜΑΤΩΝ**

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελούν αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλουν οποιασδήποτε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχουν έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο Δηλών

Μυλωνάς Γεώργιος

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

The Declarant

Mylonas Georgios

Ευχαριστίες ή Σχόλια

Θα ήθελα να ευχαριστήσω τον αναπληρωτή καθηγητή κύριο Σταμπουλή Γεώργιο για το κίνητρο και τους στόχους που έθεσε για την παρούσα διπλωματική εργασία, την αναπληρώτρια καθηγήτρια κυρία Δασκαλοπούλου Ασπασία και τον καθηγητή κύριο Βασιλακόπουλο για τις διορθώσεις και την συμμετοχή τους αξιολόγησης της εργασίας και τον υποψήφιο διδάκτορα κύριο Παπαευσταθίου Νικόλαο για τις συμβουλές και την υποστήριξη του.

Στην οικογένεια,
τους φίλους και τις φίλες μου
που με στήριξαν και με στηρίζουν.

ΑΝΑΠΤΥΞΗ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

ΜΕΤΑΔΑΤΑ ΕΠΙΣΤΗΜΟΝΙΚΩΝ ΔΗΜΟΣΙΕΥΣΕΩΝ

Μυλωνάς Γεώργιος

Περίληψη

Οι εμπειρικές μελέτες των δικτύων γνώσης διερευνούν τα μοτίβα επιστημονικών συνεργασιών σε διάφορα τεχνολογικά πεδία. Οι συγκεκριμένες μελέτες βασίζονται στην ανάλυση μεταδεδομένων επιστημονικών δημοσιεύσεων, καθώς αυτές ως προϊόντα επιστημονικής συν-συγγραφής αντικατοπτρίζουν τις αλληλοεπιδράσεις που σχηματίζονται μεταξύ ερευνητών. Παρά τον μεγάλο αριθμό ψηφιακών εργαλείων που έχουν αναπτυχθεί για την στατιστική ανάλυση μεταδεδομένων από επιστημονικές δημοσιεύσεις, παρατηρείται πως εκλείπει από το σύνολο αυτών των εργαλείων ένας αυτοματοποιημένος μηχανισμός που να αντλεί, να αποθηκεύει και να διαχειρίζεται κατάλληλα τα μεταδεδομένα αυτού του τύπου. Σε αυτή τη διπλωματική εργασία παρουσιάζεται η ανάπτυξη μιας διαδικτυακής πλατφόρμας που επιτρέπει στους χρήστες να συλλέγουν εύκολα και γρήγορα μεταδεδομένα επιστημονικών δημοσιεύσεων, τα οποία οργανώνονται σε μια ενιαία σχεσιακή δομή πίνακα, κατάλληλη για στατιστική επεξεργασία.

Λέξεις-κλειδιά: Δίκτυα γνώσης, μεταδεδομένα επιστημονικών δημοσιεύσεων, διαδικτυακή πλατφόρμα, αυτόματη άντληση δεδομένων, καθαρισμός δεδομένων, σχεδιασμός σχήματος βάσης δεδομένων, ανάπτυξη API, σχεδιασμός UI, RDBMS

DATABASE DEVELOPMENT FOR THE STATISTIC ANALYSIS OF SCIENTIFIC PUBLICATIONS METADATA

Mylonas Georgios

Abstract

The empirical studies aim to provide an understanding of patterns that are observed in the creation of connections within scientific communities. This type of studies is based on the analysis of scientific publications metadata since they reflect the interactions between researchers as co – authorship’s products. Despite the numerous digital tools that have developed for the statistical analysis of published researches metadata, it appears that an automated mechanism able to retrieve, store and manipulate this type of metadata is absent from this set of tools. In the current diploma thesis, the development of an online platform is presented. This platform will allow to its users to collect fast and with ease scientific publications metadata, which will be organized in a single relation table structure, suitable for statistic processing.

Keywords: knowledge networks, scientific publications metadata, online platform, automatic metadata retrieving, data cleaning, database schema design, API development, UI design, RDBMS

Κεφάλαιο 1 Περιεχόμενα

<i>Ευχαριστίες ή Σχόλια.....</i>	<i>xiii</i>
<i>Περίληψη.....</i>	<i>xv</i>
<i>Abstract</i>	<i>xvii</i>
<i>Κατάλογος Εικόνων.....</i>	<i>xxiii</i>
<i>Κατάλογος διαγραμμάτων</i>	<i>xxvii</i>
<i>Κατάλογος πινάκων</i>	<i>xxix</i>
<i>Συντομογραφίες.....</i>	<i>xxxi</i>
<i>Κεφάλαιο 1 ΕΙΣΑΓΩΓΗ.....</i>	<i>1</i>
1.2 Σημασία του θέματος.....	1
1.3 Επιμέρους στόχοι της διπλωματικής εργασίας.....	2
1.4 Συνοπτική περιγραφή ερευνητικών προσεγγίσεων του θέματος στην υπάρχουσα βιβλιογραφία	3
1.5 Ερευνητική και μεθοδολογική προσέγγιση	3
1.6 Διάρθρωση της διπλωματική εργασίας.....	4
<i>Κεφάλαιο 2 ΔΙΕΡΕΥΝΗΣΗ ΕΡΓΑΛΕΙΩΝ ΔΙΑΧΕΙΡΙΣΗΣ METADATA ΕΠΙΣΤΗΜΟΝΙΚΗΣ ΒΙΒΛΙΟΓΡΑΦΙΑΣ</i>	<i>5</i>
2.1 Διαστάσεις του θέματος – προβλήματος.....	5
2.2 Περιγραφή εργαλείων στην υπάρχουσα βιβλιογραφία	5
2.2.1 Conflate.....	6
2.2.2 Proof of Bibliometrics.....	7
2.2.3 Bibliometrix	8
2.2.3 VOSviewer.....	9
2.2.5 CiteNetExplorer	10
2.2.6 SciMAT	11
2.2.7 Sci2 Tool.....	12
2.2.9 Bibexcel	15
2.2.10 Vantage Point.....	16
2.2.11 ScientoPy	18

2.2.12 Pysurveillance	19
2.3 Επισκόπηση εργαλείων.....	20
2.4 Συμπεράσματα από την ανασκόπηση σχετικών ερευνών και εργαλείων	25
Κεφάλαιο 3 ΑΝΑΠΤΥΞΗ ΜΕΘΟΔΟΛΟΓΙΑΣ	29
3.1 Ανάλυση μεθοδολογίας σε επιμέρους στάδια.....	29
3.2 Ανάπτυξη βάσης δεδομένων	30
3.2.1 Επιλογή RDBMS	30
3.2.2 Περιγραφή της δομής της βάσης δεδομένων	30
3.3 Άντληση metadata επιστημονικής βιβλιογραφίας	34
3.3.1 Διερεύνηση διαθέσιμων πηγών δεδομένων	34
3.3.2 Επιλογή πηγής δεδομένων	36
3.3.3 Εργαλεία που χρησιμοποιήθηκαν	36
3.3.4 Άντληση metadata – Στάδιο 1.....	36
3.3.5 Άντληση metadata – Στάδιο 2.....	39
3.4 Backend ανάπτυξη και σχεδιασμός API.....	42
3.4.1 Εργαλεία που χρησιμοποιήθηκαν	42
3.4.2 Καταχώρηση metadata στη βάση δεδομένων	42
3.4.3 Αναγνώριση διπλοτύπων	47
3.4.4 Επιστροφή αποτελεσμάτων αναζήτησης	50
3.5 Frontend ανάπτυξη και σχεδιασμός UI	53
3.5.1 Εργαλεία που χρησιμοποιήθηκαν	53
3.5.2 Εισαγωγή και αποστολή στοιχείων αναζήτησης.....	53
3.5.3 Προβολή και διαχείριση αποτελεσμάτων αναζήτησης	54
3.6 Σύνοψη κεφαλαίου.....	56
Κεφάλαιο 4 ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ.....	59
4.1 Εισαγωγή.....	59
4.2 Παρουσίαση UI.....	59
4.2.1 Αρχική σελίδα	59
4.2.2 Σελίδα αναζήτησης	60
4.3 Σενάρια αναζήτησης.....	60
4.3.1 Σενάριο αναζήτησης 1 – Δοκιμαστική αναζήτηση.....	60
4.3.2 Σενάριο αναζήτησης 2 – Εκτενής αναζήτηση	66

4.3.3 Σενάριο αναζήτησης 3 – Εκτίμηση πλήθους δημοσιεύσεων.....	69
4.4 Σύνοψη κεφαλαίου.....	70
Κεφάλαιο 5 ΣΥΜΠΕΡΑΣΜΑΤΑ	73
5.1 Σύνοψη και συμπεράσματα.....	73
5.2 Μελλοντικές επεκτάσεις.....	74
Βιβλιογραφία.....	75
ΠΑΡΑΡΤΗΜΑΤΑ.....	81
Παράρτημα Α.....	83
Ευρετήριο Ορολογίας.....	83
Παράρτημα Β	85
Κατάλογος φακέλων και αρχείων στο Git Hub repository	85

Κατάλογος Εικόνων

Εικόνα 2.1: Ροή διεργασιών για την ανάπτυξη της πλατφόρμας Conflate	6
Εικόνα 2.2: Ροή διεργασιών για την ανάπτυξη της DLT εφαρμογής Proof of Bibliometric Indicators	7
Εικόνα 2.3: Το UI του Bibliometrix	8
Εικόνα 2.4: Στιγμιότυπο οθόνης από το κύριο παράθυρο του VOSviewer	9
Εικόνα 2.5: Το UI του CitNetExplorer και η γραφιστική απεικόνιση δικτύων που προσφέρει)	10
Εικόνα 2.6: Απεικόνιση δικτύων γνώσης μέσω του UI του SciMAT	11
Εικόνα 2.7: Οθόνη ανάλυσης στατιστικών του SciMAT	12
Εικόνα 2.8: Παράθεση στατιστικών μέσω του UI του Sci2 Tool για διαφορετικά επιστημονικά πεδία της Web of Science	13
Εικόνα 2.9: Δίκτυο συν-παραπομπών στο Sci2 Tool	13
Εικόνα 2.10: Γραφιστική αναπαράσταση δικτύου συν-παραπομπών από το CiteSpace II	15
Εικόνα 2.11: Στιγμιότυπο από το UI του Vantage Point	17
Εικόνα 2.12: Παράθυρο ανάλυσης στατιστικών στο UI του ScientoPy	19
Εικόνα 2.13: Αλληλοεπίδραση χρηστών με την πλατφόρμα της Scopus μέσω του Pysurveillance	20
Εικόνα 3.1: Διάγραμμα ροής διεργασιών που ακολουθήθηκαν	29
Εικόνα 3.2: Σχεδιασμός διαγράμματος συσχέτισης οντοτήτων μέσω ER2SQL	31
Εικόνα 3.3: Ορισμός μεθόδου getDois και αρχικοποίηση μεταβλητών	37
Εικόνα 3.4: Σχηματισμός HTTP query για την ανάκτηση αναγνωριστικών DOI	38
Εικόνα 3.5: Python κλάση για την διαχείριση metadata δημοσιεύσεων	40
Εικόνα 3.6: Python κλάση για την διαχείριση metadata συγγραφέων	40
Εικόνα 3.7: Python κλάση για την διαχείριση metadata οργανισμών	40
Εικόνα 3.8: Ανάκτηση κριτηρίων αναζήτησης και έναρξη διαδικασίας εξαγωγής metadata	43
Εικόνα 3.9: Ανάκτηση μεγεθών στηλών που περιέχουν strings	43

Εικόνα 3.10: Άντληση metadata δημοσιεύσεων και εισαγωγή τους στη βάση δεδομένων	44
Εικόνα 3.11: Διαχείριση σφαλμάτων κατά την καταχώρηση των metadata στο RDBMS..	44
Εικόνα 3.12: Άντληση metadata συγγραφέων και εισαγωγή τους στη βάση δεδομένων.	45
Εικόνα 3.13: Άντληση metadata οργανισμών και εισαγωγή τους στη βάση δεδομένων..	46
Εικόνα 3.14: Εισαγωγή IDs δημοσιεύσεων, συγγραφέων και οργανισμών στους πίνακες συσχέτισεων	47
Εικόνα 3.15: Τμήμα του SQL query όπου καθορίζονται τα επιθυμητά metadata (στήλες) για την ανάκτηση τους από την βάση	50
Εικόνα 3.16: Σχηματισμός SQL query για την εισαγωγή συνθηκών προκειμένου να ανακτηθούν οι κατάλληλες καταχωρήσεις	51
Εικόνα 3.17: Αντικείμενο απόκρισης από την διαδικασία αναζήτησης	51
Εικόνα 3.18: Έλεγχος για ύπαρξη πιθανών αντιγράφων δημοσιεύσεων εντός των δεδομένων που επιστρέφονται	52
Εικόνα 3.19: Αποστολή των παραμέτρων εισόδου στο API μέσω HTTP GET μεθόδου	54
Εικόνα 3.20: Μέθοδος για την συμπλήρωση της λίστας αντιγράφων προς διαγραφή	55
Εικόνα 3.21: Μέθοδος για την εκκαθάριση δεδομένων από διπλότυπα.....	56
Εικόνα 3.22: Αρχιτεκτονική της πλατφόρμας	58
Εικόνα 4.1: Αρχική σελίδα της πλατφόρμας Intelnetus.....	59
Εικόνα 4.2: Σελίδα αναζήτησης της πλατφόρμας Intelnetus.....	60
Εικόνα 4.3: Παράμετροι εισόδου Σεναρίου Αναζήτησης 1	61
Εικόνα 4.4: Αναμενόμενος αριθμός διασυνδέσεων δημοσιεύσεων και συγγραφέων.....	62
Εικόνα 4.5: Αναμενόμενος αριθμός διασυνδέσεων δημοσιεύσεων και οργανισμών.....	62
Εικόνα 4.6: Πραγματικός αριθμός διασυνδέσεων δημοσιεύσεων και συγγραφέων	62
Εικόνα 4.7: Πραγματικός αριθμός διασυνδέσεων δημοσιεύσεων και οργανισμών	62
Εικόνα 4.8: Τμήμα πίνακα συσχέτισης με τα metadata δημοσιεύσεων	63
Εικόνα 4.9: Τμήμα πίνακα συσχέτισης με τα metadata profiles συγγραφέων	63
Εικόνα 4.10: Τμήμα πίνακα συσχέτισης με τα metadata οργανισμών	64
Εικόνα 4.11: Menu διαχείρισης διπλοτύπων για το Σενάριο Αναζήτησης 1	65
Εικόνα 4.12: Επιλογή διπλοτύπων προς διαγραφή για το Σενάριο Αναζήτησης 1	65
Εικόνα 4.13: Δείγμα αρχείου εξόδου Excel με τα τελικά metadata	66
Εικόνα 4.14: Δείγμα αρχείου εξόδου CSV με τα τελικά metadata	66

Εικόνα 4.15: Αναμενόμενο πλήθος διασυνδέσεων άρθρων – συγγραφέων	67
Εικόνα 4.16: Πραγματικό πλήθος διασυνδέσεων άρθρων – συγγραφέων.....	67
Εικόνα 4.17: Αναμενόμενο πλήθος διασυνδέσεων άρθρων – οργανισμών	67
Εικόνα 4.18: Πραγματικό πλήθος διασυνδέσεων άρθρων – οργανισμών.....	67
Εικόνα 4.19: Menu διαχείρισης διπλοτύπων για το Σενάριο Αναζήτησης 2.....	68
Εικόνα 4.20: Πιθανά διπλότυπα συγγραφέων για το Σενάριο Αναζήτησης 2.....	68
Εικόνα 4.21: Πιθανά διπλότυπα συγγραφέων για το Σενάριο Αναζήτησης 2.....	68
Εικόνα 4.22: Αναλυτική παράθεση αριθμού δημοσιεύσεων ανά έτος για το Σενάριο Αναζήτησης 3.....	69

Κατάλογος διαγραμμάτων

Διάγραμμα 2.1 Κάλυψη των λειτουργιών από εργαλεία επεξεργασίας δεδομένων επιστημονικής βιβλιογραφίας	27
Διάγραμμα 4.1 Αριθμός δημοσιεύσεων ανά έτος για το Σενάριο Αναζήτησης 3	

Κατάλογος πινάκων

Πίνακας 2.1 Επισκόπηση των βιβλιομετρικών εργαλείων που εξετάσθηκαν στην υποενότητα	21
Πίνακας 3.1 Διαθέσιμα metadata δημοσιεύσεων από πλατφόρμες επιστημονικής βιβλιογραφίας	35
Πίνακας 3.2 Διαθέσιμα metadata συγγραφέων από πλατφόρμες επιστημονικής βιβλιογραφίας	35
Πίνακας 3.3 Διαθέσιμα metadata οργανισμών από πλατφόρμες επιστημονικής βιβλιογραφίας	35
Πίνακας 4.1 Τεχνικές λεπτομέρειες σεναρίων αναζήτησης	71

Συντομογραφίες

AGR	Average Growth Rate
ADY	Average Documents Per Year
API	Application Program Interface
CSV	Comma Separated Values
DLT	Distribution Ledger Technology
DOI	Digital Object Identifier
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
JSON	JavaScript Object Notation
PDLY	Percentage of Documents in Last Years
RDBMS	Relational Database Management System
SQL	Syntax Query Language
UI	User Interface
κ.ά.	και άλλοι
κ.ο.κ	και ούτω καθεξής
π.χ.	παραδείγματος χάριν

Κεφάλαιο 1 ΕΙΣΑΓΩΓΗ

1.1 Πρόκληση και σκοπός της διπλωματικής εργασίας

Η πρόκληση στην οποία επιδιώκει να ανταποκριθεί η διπλωματική εργασία είναι το πως μπορούν οι χρήστες διαδικτυακών βάσεων δεδομένων, οι οποίες εμπεριέχουν επιστημονικές δημοσιεύσεις, να αντλούν πιο εύκολα και γρήγορα μεταδεδομένα (metadata) επιστημονικών δημοσιεύσεων. Σκοπός της διπλωματικής εργασίας είναι η ανάπτυξη ενός αυτοματοποιημένου μηχανισμού για πιο εύκολη και γρήγορη άντληση metadata επιστημονικών δημοσιεύσεων από διαδικτυακές βάσεις δεδομένων σε ενιαία βάση (πλατφόρμα) για την ευχερή επεξεργασία. Συγκριτικά με παρόμοια εργαλεία που υπάρχουν, η πλατφόρμα αυτής της διπλωματικής εργασίας παρέχει τη δυνατότητα άντλησης των πληροφοριών επιστημονικών άρθρων και των οντοτήτων που σχετίζονται με τη συγγραφή τους (συγγραφείς, οργανισμοί) και η αυτοματοποιημένη οργάνωση τους σε μια ενιαία σχεσιακή δομή πίνακα.

1.2 Σημασία του θέματος

Η ανάλυση βιβλιομετρικών δεδομένων είναι χρήσιμη για τη μελέτη της δυναμικής της επιστημονικής έρευνας [1],[2]. Πρόσφατα έχει αναπτυχθεί η βιβλιογραφία για τη μελέτη των δικτύων γνώσης μέσα από τα βιβλιομετρικά δεδομένα [3],[4],[5], τα οποία συνήθως προκύπτουν από τα μεταδεδομένα των δημοσιεύσεων ή άλλων στοιχείων καινοτομίας όπως οι πατέντες.

Οι επιστημονικές δημοσιεύσεις θεωρούνται δίκτυα συν-συγγραφής γιατί αντικατοπτρίζουν επιστημονικές συνεργασίες [6],[7],[8] που προκύπτουν μέσα από τη δικτύωση των επιστημόνων ή των ερευνητών [3],[9]. Αυτά τα δίκτυα συν-συγγραφής θεωρούνται και δίκτυα γνώσης [4],[9]. Μελετώντας τα δίκτυα γνώσης αναλύονται οι διαδικασίες διάχυσης γνώσης και δημιουργίας νέας [10]. Με τα μετα-δεδομένα από επιστημονικές δημοσιεύσεις μπορούμε να διερευνήσουμε τα μοτίβα των επιστημονικών συνεργασιών [4].

Επιπλέον μπορούν να μελετηθούν οι παράγοντες που επηρεάζουν την ανάπτυξη των επιστημονικών συνεργασιών όπως η δομή του δικτύου [3],[5],[11],[12], η εγγύτητα των συγγραφέων ή των οργανισμών στους οποίους εργάζονται (γεωγραφική, θεσμική, πολιτιστική, κοινωνική, οργανωσιακή και γνωστική) [12],[13],[14] και το επιστημονικό

αντικείμενο της ερευνητικής συνεργασίας [4]. Τα μεταδεδομένα των επιστημονικών δημοσιεύσεων μπορούν να χρησιμοποιηθούν για τη μελέτη της σχέσης μεταξύ κοινωνικού κεφαλαίου και δικτύων γνώσης [15],[16],[17],[18],[19].

Επίσης, τα μεταδεδομένα των επιστημονικών δημοσιεύσεων είναι απαραίτητα για τη διερεύνηση των παραγόντων που επηρεάζουν την ερευνητική απόδοση [3],[11],[12].

Η πλατφόρμα που αναπτύσσεται στην παρούσα εργασία, διευκολύνει και επιταχύνει την διεξαγωγή βιβλιομετρικών αναλύσεων και τη μελέτη των επιστημονικών συνεργασιών σε ποικίλα επιστημονικά πεδία. Επιπρόσθετα, συμβάλλει στην διερεύνηση παραμέτρων που επηρεάζουν τη συνεργασία για την παραγωγή (συν-συγγραφή) επιστημονικών δημοσιεύσεων, ενώ συνεισφέρει και στην έρευνα του τρόπου με τον οποίο αναπτύσσονται οι επιστημονικές συνεργασίες και των αποτελεσμάτων τους σε επίπεδο καινοτομίας.

1.3 Επιμέρους στόχοι της διπλωματικής εργασίας

Οι επιμέρους στόχοι της διπλωματικής εργασίας είναι οι εξής:

- Επιλογή RDBMS και καθορισμός του σχήματος της βάσης δεδομένων που θα εξυπηρετεί τους σκοπούς της διπλωματικής εργασίας (π.χ. αριθμός πινάκων, χαρακτηριστικά πινάκων, σχέσεις μεταξύ πινάκων, καθορισμός primary keys κτλ.).
- Ανάπτυξη αλγορίθμων και υλοποίηση μεθόδων σε κώδικα προγραμματισμού Python για την εξαγωγή πρωτογενών metadata ακαδημαϊκής βιβλιογραφίας από κοινόχρηστες βάσεις δεδομένων μέσω των αντίστοιχων APIs και την κατάλληλη προ-επεξεργασία των εξαγόμενων δεδομένων.
- Σχεδιασμός και ανάπτυξη API μέσω της Python βιβλιοθήκης Flask που θα πραγματοποιεί αποτελεσματικά :
 - Την λήψη κριτηρίων αναζήτησης από τους χρήστες μέσω HTTP αιτημάτων.
 - Την άντληση και την προ-επεξεργασία metadata μέσω των κατάλληλων μεθόδων που έχουν αναπτυχθεί.
 - Την καταχώρηση των εξαγόμενων δεδομένων στο σχήμα βάσης που φιλοξενείται στο RDBMS.
 - Τον εντοπισμό πιθανών αντιγράφων για κάθε καταχώρηση δημοσίευσης, ερευνητή ή οργανισμού.

- Την κατάλληλη μορφοποίηση και την επιστροφή των αποτελεσμάτων στους χρήστες.
- Ανάπτυξη ενός απλού στην χρήση UI που θα επιτρέπει στους χρήστες:
 - Την επιλογή παραμέτρων αναζήτησης.
 - Την προβολή και την διαχείριση των αποτελεσμάτων αναζήτησης.

1.4 Συνοπτική περιγραφή ερευνητικών προσεγγίσεων του θέματος στην υπάρχουσα βιβλιογραφία

Μελετώντας έρευνες στην υπάρχουσα βιβλιογραφία που να σχετίζονται με τον σχεδιασμό και την ανάπτυξη, διαπιστώνεται πως οι ερευνητικές προσεγγίσεις που προτείνονται συγκλίνουν στην διαμόρφωση τριών βασικών ζητημάτων: άντληση, επεξεργασία και παράθεση metadata στους χρήστες προκειμένο να αξιοποιηθούν όπως επιθυμείται σε κάθε περίπτωση.

Η διαδικασία της άντλησης δεδομένων πραγματοποιείται στην πλειονότητα των περιπτώσεων χειροκίνητα μέσω της δημιουργίας αρχείων εξόδου από τις πλατφόρμες επιστημονικού υλικού που αξιοποιούνται σε κάθε μελέτη περίπτωσης. Σπανιότερα, αυτό επιτυγχάνεται μέσω κώδικα που αναπτύσσουν οι ίδιοι οι συγγραφείς και εντάσσουν στο εύρος λειτουργίας των αντίστοιχων λογισμικών.

Στο στάδιο της επεξεργασίας των metadata δημοσιεύσεων, αναπτύσσονται αλγόριθμοι κυρίως για την ομογενοποίηση κειμένου και την αξιοποίηση αυτού για τον εντοπισμό πιθανών αντιγράφων. Η διαχείριση (απαλοιφή ή συγχώνευση) των αντιγράφων είτε επιτελείται από τους ίδιους τους αλγόριθμους είτε παραδίδεται στους χρήστες για μεγαλύτερη ευελιξία.

Η παράθεση των τελικών δεδομένων στους χρήστες αφορά αφενός την προβολή αυτών μέσω κατάλληλων δομών (π.χ. πίνακες, γραφήματα, διαγράμματα) και αφετέρου την περαιτέρω επεξεργασία τους εκ μέρους τους, μέσω δυνατοτήτων όπως φιλτράρισμα καταγραφών, διαχείριση διπλοτύπων ή εξαγωγή δεδομένων μέσω αρχείων εξόδου.

1.5 Ερευνητική και μεθοδολογική προσέγγιση

Αρχικά επιλέγεται το RDBMS στο οποίο θα φιλοξενηθεί η βάση δεδομένων της πλατφόρμας, βάσει κριτηρίων ευελιξίας και απόδοσης διαχείρισης δεδομένων, ενώ καθορίζεται το σχήμα της βάσης δεδομένων. Στη συνέχεια, γίνεται επιλογή των βάσεων

δεδομένων επιστημονικών δημοσιεύσεων, λαμβάνοντας υπόψιν τις δυνατότητες και τους περιορισμούς των αντίστοιχων APIs. Έπειτα, αναπτύσσεται κώδικας σε γλώσσα Python για την άντληση και την εισαγωγή των βιβλιομετρικών δεδομένων στη βάση δεδομένων που έχει αναπτυχθεί και παρουσιάζεται η υλοποίηση αλγόριθμων ανίχνευσης διπλοεγγραφών, για την διασφάλιση της ποιότητας και της αξιοπιστίας των δεδομένων. Τέλος, πραγματοποιείται τόσο η backend ανάπτυξη για το σχεδιασμό API που λαμβάνει και διαχειρίζεται κατάλληλα αιτήματα χρηστών, όσο και η frontend ανάπτυξη για τη δημιουργία UI που επιτρέπει την εύκολη αλληλεπίδραση των χρηστών με την πλατφόρμα.

1.6 Διάρθρωση της διπλωματική εργασίας

Στη συνέχεια η διπλωματική εργασία διαρθρώνεται ως εξής:

Στο Δεύτερο Κεφάλαιο παρουσιάζονται οι δημοσιευμένες έρευνες που σχετίζονται την μεθοδολογική προσέγγιση που ακολουθείται στην παρούσα διπλωματική εργασία.

Στο Τρίτο Κεφάλαιο αναπτύσσεται τη μεθοδολογία σχεδιασμού και ανάπτυξης της πλατφόρμας.

Στο Τέταρτο Κεφάλαιο παρουσιάζεται η διεπαφή χρήστη της πλατφόρμας καθώς τα αποτελέσματα που επιστρέφονται έπειτα από μια σειρά σεναρίων αναζήτησης.

Τέλος, στο Πέμπτο Κεφάλαιο παραθέτω συμπεράσματα και προτάσεις για μελλοντικές επεκτάσεις.

Κεφάλαιο 2 ΔΙΕΡΕΥΝΗΣΗ ΕΡΓΑΛΕΙΩΝ ΔΙΑΧΕΙΡΙΣΗΣ METADATA ΕΠΙΣΤΗΜΟΝΙΚΗΣ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

2.1 Διαστάσεις του θέματος – προβλήματος

Σε αυτό το κεφάλαιο γίνεται αναφορά υπαρχόντων εργαλείων που βρίσκουν εφαρμογή στην εκπόνηση μελετών και αξιολογήσεων επιστημονικής βιβλιογραφίας, καθώς και των αντίστοιχων δημοσιεύσεων στις οποίες πραγματοποιείται περιγραφή και αξιολόγηση αυτών των εργαλείων.

Σκοπός του συγκεκριμένου κεφαλαίου είναι η εξαγωγή συμπερασμάτων σχετικά με το ποιες πτυχές της εκπόνησης επιστημονικομετρικών ερευνών κατέχουν μεγαλύτερη ή μικρότερη κάλυψη από τα ψηφιακά μέσα που έχουν αναπτυχθεί και αξιοποιούνται σε αυτές.

Για τον λόγο αυτό, το κύριο θέμα της διπλωματικής εργασίας αναλύεται σε επιμέρους διαστάσεις, ώστε να εξετασθεί το κατά πόσο συνεισφέρουν σε αυτές τα εργαλεία που θα αναφερθούν παρακάτω. Συγκεκριμένα, οι επί μέρους διαστάσεις που λαμβάνονται υπόψη στην παρούσα έρευνα καθώς και στις μελλοντικές επεκτάσεις της είναι:

- Παροχή code - free UI, μέσω του οποίου οι χρήστες της πλατφόρμας θα μπορούν εύκολα, γρήγορα και άμεσα να αξιοποιούν τις δυνατότητες της.
- Εντοπισμός πολλαπλών εκδόσεων δημοσιεύσεων ή profiles συγγραφέων και οργανισμών που αντιστοιχούν στην ίδια οντότητα.
- Αυτοματοποίηση της διαδικασίας της άντλησης πρωτογενών δεδομένων επιστημονικής βιβλιογραφίας.
- Δυνατότητα διαχείρισης των αποτελεσμάτων που επιστρέφονται, όπως η δημιουργία αρχείων εξόδου, φιλτράρισμα καταγραφών και η διαχείριση αντιγράφων.
- Αυτόματη δημιουργία δομών συσχέτισης αυξημένης ακρίβειας.

2.2 Περιγραφή εργαλείων στην υπάρχουσα βιβλιογραφία

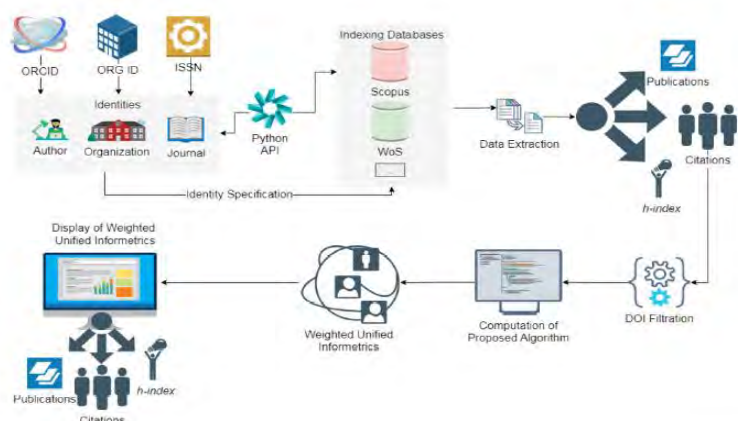
Η ανάπτυξη ψηφιακών εργαλείων για την άντληση, επεξεργασία και ανάλυση μεταδεδομένων επιστημονικής βιβλιογραφίας, ώστε να αξιοποιηθούν για την μελέτη δικτύων γνώσης και των δυναμικών τους, έχει υπάρξει αντικείμενο μελέτης πολλαπλών ερευνών ανά τα έτη, με τις πιο άμεσα συ-σχετιζόμενες εξ αυτών με την πλατφόρμα που

αναπτύσσεται στη συγκεκριμένη διπλωματική εργασία, να εξετάζονται στις παραγράφους που ακολουθούν.

2.2.1 Conflate

Πρόσφατα, οι Khurana κ.ά. (2022) [21] πρότειναν την ενοποίηση των μετρικών συγγραφέων από διαφορετικές βάσεις δεδομένων σε μια ενιαία πλατφόρμα. Συγκεκριμένα οι ερευνητές προτείνουν έναν αλγόριθμο σταθμισμένου υπολογισμού των μετρικών συγγραφέων (π.χ. αριθμός αναφορών από άλλα άρθρα, h-index συγγραφέων) από τα αντίστοιχα μετρικά των βάσεων δεδομένων των Scopus και Web of Science. Τα εκ νέου υπολογισμένα μετρικά αποθηκεύονται στην πλατφόρμα Conflate, ώστε να αξιοποιηθούν για την μελέτη της επιρροής ερευνητών, οργανισμών και επιστημονικών περιοδικών και μάλιστα σε σύγκριση με τα αντίστοιχα μετρικά των Scopus και Web of Science.

Στην Εικόνα 2.1 περιγράφεται η ροή διεργασιών ώστε να αναπτυχθεί η συγκεκριμένη πλατφόρμα. Στο πρώτο στάδιο, πραγματοποιείται η άντληση μεταδεδομένων δημοσιεύσεων, συγγραφέων και οργανισμών, όπως ο αριθμός παραπομπών δημοσιεύσεων και ο h-Index συγγραφέων, από τις βάσεις δεδομένων των Scopus και Web of Science. Στη συνέχεια ακολουθεί το φιλτράρισμα των καταχωρήσεων μέσω της αξιοποίησης των αναγνωριστικών DOI και ORCID ID για την απαλοιφή διπλοτύπων εγγράφων και profiles ερευνητών αντίστοιχα. Τον καθαρισμό των δεδομένων ακολουθούν η ενοποίηση των βιβλιογραφικών μετρικών από τον προτεινόμενο αλγόριθμο που παρουσιάζουν οι συγγραφείς και τελικά, η αξιοποίηση αυτών των μετρικών μέσω του UI της πλατφόρμας, για την εκπόνηση επιστημονικομετρικών μελετών από τους χρήστες

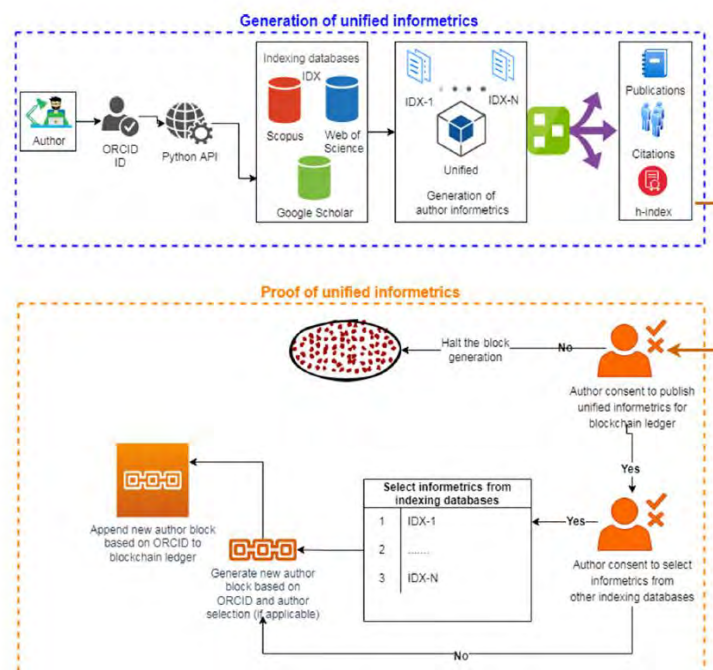


Εικόνα 2.1: Ροή διεργασιών για την ανάπτυξη της πλατφόρμας Conflate (πηγή: Khurana κ.ά., 2022, A Comparative Analysis of Unified Informetrics with Scopus and Web of Science, Journal of Scientometric Research).

2.2.2 Proof of Bibliometrics

Στην ανάπτυξη της πλατφόρμας Conflate βασίστηκαν ξανά οι Khurana κ.ά. (2022) [22], για την ενοποίηση μετρικών ερευνητών από πολλαπλές βάσεις δεδομένων, σε έναν μηχανισμό DLT. Σκοπός της συγκεκριμένης έρευνας είναι η ανάπτυξη μιας αποκεντρωμένης εφαρμογής κοινής συναίνεσης με όνομα Proof of Bibliometric Indicators στην οποία αποθηκεύονται ενοποιημένα μετρικά και στοιχεία συγγραφέων ως blocks σε ένα καθολικό blockchain, αφού έχουν ήδη αντληθεί από τις βάσεις δεδομένων των Scopus, Web of Science και Google Scholar. Τα στοιχεία αυτά στη συνέχεια μπορούν να αξιοποιηθούν από ερευνητές και οργανισμούς για την εκπόνηση επιστημονικομετρικών μελετών.

Η ροή διεργασιών που ακολουθήθηκε για την δημιουργία της συγκεκριμένης εφαρμογής απεικονίζεται στην Εικόνα 2.2. Όπως παρατηρείται, παρουσιάζει σημαντικές ομοιότητες με την αντίστοιχη της πλατφόρμας Conflate, με τη διαφορά ότι σε αυτήν την περίπτωση τα δεδομένα αποθηκεύονται στο καθολικό blockchain της DLT εφαρμογής αντί μιας σχεσιακής βάσης δεδομένων.



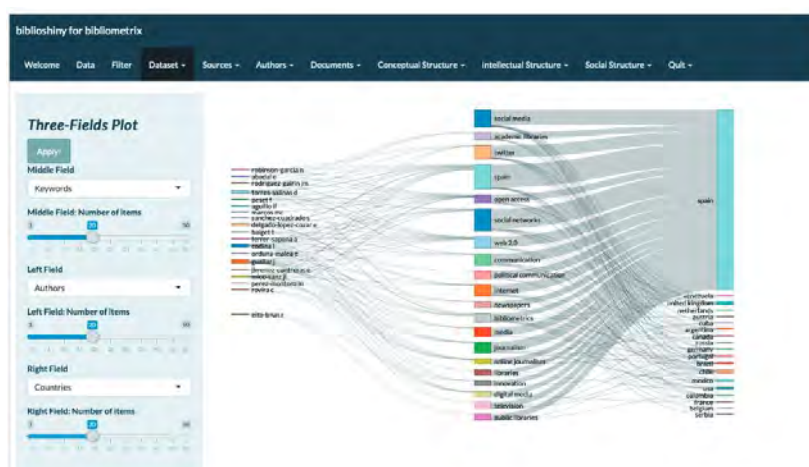
Εικόνα 2.2: Ροή διεργασιών για την ανάπτυξη της DLT εφαρμογής Proof of Bibliometric Indicators (πηγή: Khurana, 2022, Proof of Bibliometric Indicators: A blockchain based consensus protocol for publications, Multimedia Tools and Applications).

2.2.3 Bibliometrix

Η δουλειά που παρουσιάστηκε από τους Aria & Cuccurullo (2017) [23] αφορά την ανάπτυξη του πακέτου Bibliometrix σε γλώσσα προγραμματισμού R, που προσφέρει ένα σύνολο συναρτήσεων για την πραγματοποίηση επιστημονικομετρικών αναλύσεων. Στο πρώτο στάδιο της προτεινόμενης ροής εργασιών, επιτελείται η συλλογή δεδομένων από τις βάσεις δεδομένων των Scopus και Web of Science μέσω των αντίστοιχων APIs και έπειτα η οργάνωση τους σε dataframes. Στο δεύτερο στάδιο πραγματοποιείται ο υπολογισμός των μετρικών (π.χ αριθμός συν-συγγραφών, συν-παραθέσεων) που χαρακτηρίζουν το εκάστοτε δίκτυο ερευνητών, ενώ στο τρίτο και τελευταίο στάδιο, λαμβάνει χώρα η οπτικοποίηση του ερευνητικού δικτύου μέσω των μετρικών που παράχθηκαν στο δεύτερο στάδιο.

Ο Dervis (2019) [24] παρουσιάζει ως σημαντικό πλεονέκτημα του εργαλείου το UI που διαθέτει (Εικόνα 2.3), το οποίο καθιστά την διαδικασία επεξεργασίας δεδομένων και ανάλυσης στατιστικών αποτελεσμάτων πιο απλή και εύκολη για χρήστες μη εξοικειωμένους με ανάπτυξη κώδικα R. Ο ίδιος προσθέτει στο άρθρο του πως το Bibliometrix παρέχει και μεγαλύτερη ευελιξία στην ανάλυση των στατιστικών αποτελεσμάτων συγκριτικά με παρόμοιου τύπου εργαλεία, καθώς αυτή μπορεί να πραγματοποιηθεί σε πολλαπλά επίπεδα (εγγράφων, λέξεων κ.ά.), ανάλογα με τις ανάγκες της εκάστοτε έρευνας.

Παρόλο που το συγκεκριμένο εργαλείο παρέχει χρήσιμες συναρτήσεις για τη δημιουργία και την απεικόνιση δικτύων γνώσης και την εξαγωγή στατιστικών στοιχείων, εκλείπουν εξειδικευμένες ρουτίνες καθαρισμού δεδομένων, πέραν κάποιων βασικών συναρτήσεων που αφορούν κυρίως διορθώσεις σε επίπεδο κειμένου και όχι εγκυρότητας, όπως επισημαίνουν οι Aria & Cuccurullo (2017) [23].

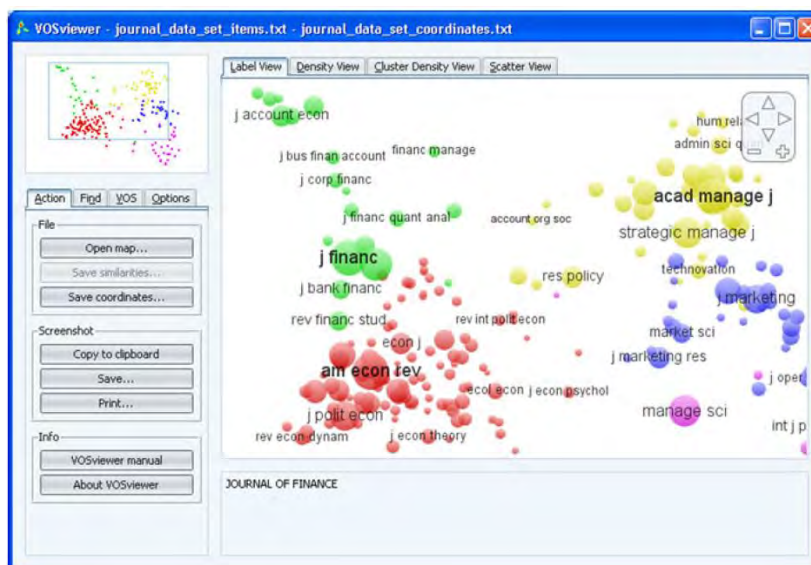


Εικόνα 2.3: Το UI του Bibliometrix (πηγή: Moral-Munoz κ.ά., 2020, Software tools for conducting bibliometric analysis in science: An up-to-date review, El profesional de la información).

2.2.3 VOSviewer

Οι Eck κ.ά. (2009) [25] παρουσιάζουν στο άρθρο τους το πρόγραμμα VOSviewer, το οποίο εξειδικεύεται στην κατασκευή και την απεικόνιση βιβλιομετρικών χαρτών. Η εκτενής περιγραφή της λειτουργίας του στο άρθρο περιέχει την άντληση δεδομένων από τους καταλόγους επιστημονικών κατηγοριών της Thompson Reuters και την αξιοποίησή τους για τον σχεδιασμό δικτύων που χαρτογραφούν την αλληλεπίδραση μεταξύ επιστημονικών αρθρογράφων, περιοδικών και οργανισμών.

Σύμφωνα με τους συγγραφείς, η ικανότητα του VOSviewer να απεικονίζει αποτελεσματικά ογκώδεις χάρτες επιστημονικών δικτύων αποτελεί και το μεγαλύτερο πλεονέκτημα του εργαλείου, όπως επισημαίνεται και από τους Cobo κ.ά. (2011) [26] και Mogali (2012) [27]. Το εργαλείο δέχεται ως είσοδο πίνακες συσχέτισης από τους οποίους εξάγει μετρικές των αλληλεπιδράσεων και των συσχετίσεων μεταξύ των εκάστοτε οντοτήτων που λαμβάνονται υπόψιν, ώστε τελικά να απεικονίσει την εξαγόμενη πληροφορία υπό τη μορφή δικτύων γνώσης (Εικόνα 2.4). Η έμφαση που δίνεται στην ευκρινή και αποτελεσματική γραφιστική αναπαράσταση των επιστημονικομετρικών χαρτών, γίνεται αντιληπτή μέσω της παροχής δυνατοτήτων μεγέθυνσης, αναζήτησης και αυτόματης μορφοποίησης των κόμβων που συνιστούν τα εξαγόμενα δίκτυα.

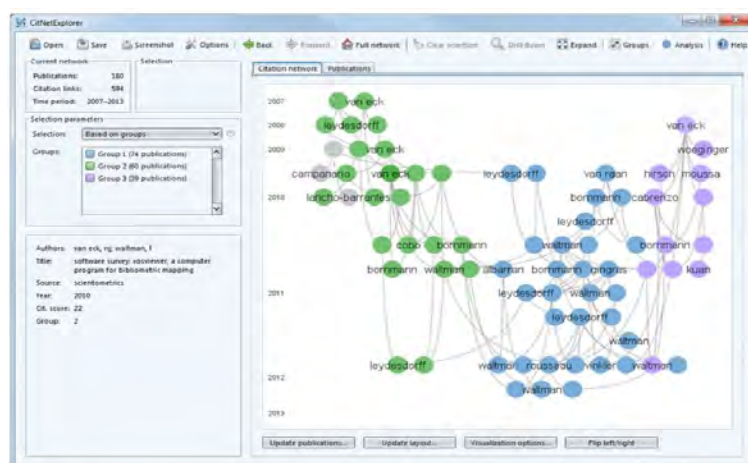


Εικόνα 2.4: Στιγμιότυπο οθόνης από το κύριο παράθυρο του VOSviewer (πηγή: Eck & Waltman, 2010, Software survey: VOSviewer, a computer program, Scientometrics).

2.2.5 CiteNetExplorer

Το VOSviewer αποτέλεσε τη βάση για την ανάπτυξη από τους Eck κ.ά. (2014) [28] του λογισμικού CitNetExplorer, ενός ανεπτυγμένου σε Java εργαλείου που παρέχει τη δυνατότητα ανάλυσης και οπτικοποίησης δικτύων επιστημονικών αναφορών. Πολλαπλές μελέτες περιπτώσεων εξετάζονται μάλιστα εντός του άρθρου, με στόχο την ανάδειξη της αποτελεσματικότητας του εργαλείου να διαχειρίζεται δίκτυα εκατομμυρίων καταχωρήσεων, όπως επισημαίνεται από τους συγγραφείς. Αυτές οι μελέτες περιπτώσεων βασίζονται στη χειροκίνητη άντληση δεδομένων από την πλατφόρμα της Web of Science και στην αλληλεπίδραση του χρήστη με την εξαγόμενη πληροφορία μέσω του UI του εργαλείου που παρουσιάζεται στην Εικόνα 2.5, κάτι που καθιστά τη διαδικασία πιο κατανοητή σε χρήστες μη εξοικειωμένους με ανάπτυξη κώδικα προγραμματισμού.

Το CitNetExplorer χρησιμοποιεί ένα σύνολο από βιβλιομετρικά χαρακτηριστικά για την περιγραφή των δημοσιεύσεων και την εκτίμηση του αντίκτυπου που έχει η καθεμία ως προϊόν επιστημονικής συν-συγγραφής. Σε αυτά τα χαρακτηριστικά περιλαμβάνονται μεταξύ άλλων, μεταδεδομένα όπως τίτλος, έτος έκδοσης, συγγραφείς και αριθμός παραπομπών ενός άρθρου από άλλες δημοσιεύσεις. Τα δεδομένα αυτά αξιοποιούνται από το λογισμικό για τη δημιουργία clusters δημοσιεύσεων, από των οποίων τη συνένωση προκύπτει και το δίκτυο συν-παραπομπών. Οι Eck & Waltman (2014) [28] αναφέρουν πως δίνεται μάλιστα η δυνατότητα στους χρήστες να αναζητήσουν στοχευμένα δημοσιεύσεις εντός του δικτύου, βάσει τίτλου, έτους δημοσίευσης, ονόματος συγγραφέα ή περιοδικού, ώστε να εστιάσουν την ανάλυση τους στα αντίστοιχα clusters που ανήκουν αυτές οι δημοσιεύσεις.



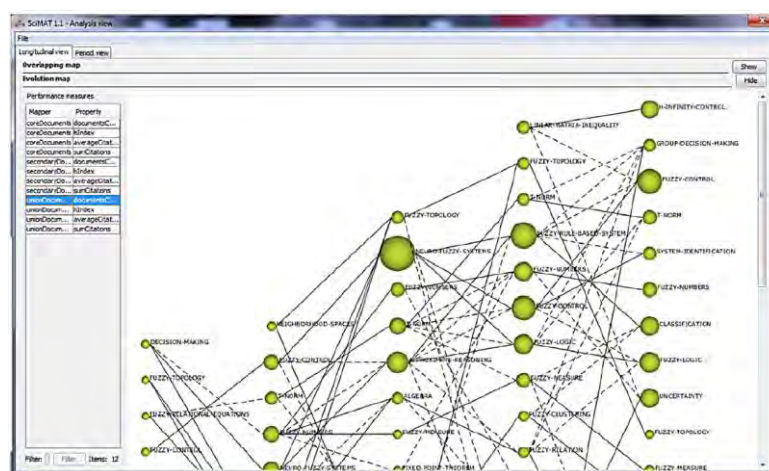
Εικόνα 2.5: Το UI του CitNetExplorer και η γραφιστική απεικόνιση δικτύων που προσφέρει (πηγή: Eck κ.ά., 2014, CiteNetExplorer: A new software tool for analyzing and visualizing citation networks, Journal of Informetrics).

2.2.6 SciMAT

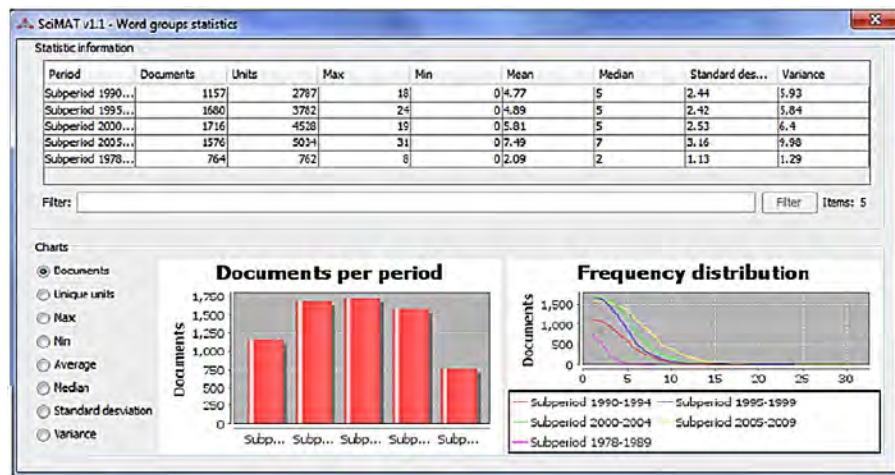
Ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη των Cobo κ.ά. (2012) [29], η οποία εστιάζει στην ανάπτυξη του ψηφιακού εργαλείου SciMAT. Το συγκεκριμένο εργαλείο σχετίζεται με την εκτέλεση επιστημονικής χαρτογράφησης μέσω των μεθόδων, των συναρτήσεων και των αλγορίθμων που εμπεριέχονται στο API του εργαλείου. Σκοπός του API είναι να επιτελεί τις διαδικασίες φόρτωσης, επεξεργασίας και οπτικοποίησης της εξαγόμενης από τα βιβλιομετρικά δεδομένα πληροφορίας.

Τα βιβλιομετρικά δεδομένα που εισάγονται στο SciMAT οργανώνονται σε επιμέρους Οντότητες (Entities) όπως αναφέρονται, που η καθεμία διασυνδέεται με ένα συγκεκριμένο σύνολο βιβλιομετρικών δεδομένων. Χαρακτηριστικά παραδείγματα Οντοτήτων αποτελούν αυτές των Δημοσιεύσεων (Papers), των Συγγραφέων (Authors) και των Συνδέσεων (Affiliations), των οποίων οι συσχετίσεις που περιγράφονται μέσω των μεταδεδομένων (π.χ. οι συγγραφείς κάθε άρθρου, οι οργανισμοί με τους οποίους συνδέεται κάθε συγγραφέας) παρέχουν όλη την απαραίτητη πληροφορία για τη δημιουργία επιστημονικών χαρτών της Εικόνας 2.6 και την εξαγωγή στατιστικών στοιχείων της Εικόνας 2.7.

Ως κύριο πλεονέκτημα του SciMAT αναδεικνύεται η ευελιξία που δίδεται στον χρήστη να επιλέγει μέσω του UI τις μετρήσεις, τους αλγορίθμους και τις τεχνικές που διεξάγονται κατά της διάρκεια της ανάλυσης δεδομένων (π.χ. h-index, μέθοδοι καθαρισμού δεδομένων) και της προβολής των στατιστικών τους. Επιπρόσθετα, παρέχει στάδιο προ-επεξεργασίας των πρωτογενών δεδομένων για την εξάλειψη πιθανών λαθών, όπως διπλότυπα ή ορθογραφικά λάθη.



Εικόνα 2.6: Απεικόνιση δικτύων γνώσης μέσω του UI του SciMAT (Cobo κ.ά., 2012, SciMAT A New Science Mapping Analysis Software Tool, Journal of the American Society for Information Science and Technology).



Εικόνα 2.7: Οθόνη ανάλυσης στατιστικών του SciMAT (πηγή: Cobo κ.ά., 2012, SciMAT A New Science Mapping Analysis Software Tool, Journal of the American Society for Information Science and Technology).

2.2.7 Sci2 Tool

Η Solomon (2015) [30] αξιολογεί στην έρευνα της την αποτελεσματικότητα του Sci2 Tool, μιας ψηφιακής εργαλειοθήκης που ειδικεύεται στην χρονική, χωρική και θεματική ανάλυση επιστημονικών δικτύων και στην οπτικοποίηση ακαδημαϊκών συνόλων δεδομένων. Η συγγραφέας συγκεκριμένα αξιοποιεί το εν λόγω εργαλείο για τη δημιουργία και την γραφιστική αναπαράσταση δικτύων γνώσης, εισάγοντας σε αυτό μεταδεδομένα επιστημονικών άρθρων (π.χ. αριθμό παραπομπών, h-index συγγραφέων και οργανισμούς χρηματοδότησης δημοσιεύσεων) που έχουν ανακτηθεί χειροκίνητα από τη βάση δεδομένων της Web of Science. Η στατιστική ανάλυση των δεδομένων πραγματοποιείται μέσω των αντίστοιχων γραφημάτων και διαγραμμάτων της Εικόνας 2.8 και κυρίως, μέσω της απεικόνισης πολλαπλών τύπων επιστημονικών δικτύων που προσφέρει το Sci2 Tool στην Εικόνα 2.9, όπως δίκτυα συν-συγγραφής, συν-παραπομπών και βιβλιογραφικής σύζευξης.

Αντίστοιχη ευελιξία παρουσιάζεται και στην ανάγνωση διαφορετικού τύπου format CSV αρχείων εξόδου, με χαρακτηριστικά παραδείγματα αυτά των Web of Science, Scopus και National Science Foundation, όπως σημειώνεται από τους Sangam & Mogali (2012) [27], ενώ οι Cobo κ.ά. (2011) [26] κάνουν αναφορά και στη δυνατότητα του Sci2 Tool να καθαρίζει τα εισαγόμενα σε αυτό δεδομένα. Όπως περιγράφεται μάλιστα στην περίπτωση μελέτης των Lewis & Alpi (2017) [31], ο καθαρισμός των δεδομένων γίνεται σε δύο στάδια. Στο πρώτο στάδιο εισάγονται τα πρωτογενή δεδομένα στο Sci2, ώστε να εντοπιστούν οι διπλοεγγραφές μέσω κατάλληλης επεξεργασίας. Στο δεύτερο στάδιο, εξάγεται από το Sci2 Tool ένα αρχείο CSV με σκοπό να επανεισαχθεί στο εργαλείο, αυτή τη φορά για την

2.2.8 CiteSpace II

Ο Chen (2005) [33] παρουσιάζει στο άρθρο του την ανάπτυξη του CiteSpace II, ενός ψηφιακού συστήματος που μπορεί να αξιοποιηθεί για την ανίχνευση και την οπτικοποίηση αναδυόμενων τάσεων και μεταβολών σε επιστημονικά πεδία συναρτήσει του χρόνου.

Σύμφωνα με τη ροή λειτουργίας που παρουσίασε ο Chen, αρχικά πρέπει να γίνει ταυτοποίηση ενός ερευνητικού πεδίου, με όσο το δυνατόν πιο γενικευμένους όρους για τη μέγιστη συμπερίληψη των βασικών υπο-πεδίων. Μετά ακολουθεί η συλλογή μεταδεδομένων βιβλιογραφικών καταχωρήσεων, όπως τίτλοι, περιλήψεις και αναφορές δημοσιεύσεων, από τις βάσεις δεδομένων των Web of Science και PubMed, βάσει των όρων που χρησιμοποιήθηκαν για την ταυτοποίηση του ερευνητικού πεδίου.

Στη συνέχεια πραγματοποιείται φιλτράρισμα των δημοσιεύσεων μέσω λέξεων ή φράσεων – κλειδιών, χρονικού πλαισίου και βιβλιογραφικών μετρικών (π.χ. αριθμός αναφορών, συν - αναφορών), ώστε να εξαχθεί από τα τελικά μεταδεδομένα η δομή του εκάστοτε επιστημονικού δικτύου. Ο χρήστης έτσι δύναται να αλληλοεπιδρά με την γραφιστική απεικόνιση του δικτύου που παρουσιάζεται στην Εικόνα 2.10 και να έχει τον έλεγχο της προβολής των χαρακτηριστικών του (π.χ. κόμβοι, ταμπέλες, περιγραφές κειμένου), με στόχο την βελτιστοποίηση της διαδικασίας της οπτικοποίησης και άρα αξιοποίησης της βιβλιογραφικής πληροφορίας.

Οι έρευνες των Cobo κ.ά. (2011) [26] και Sangam & Mogali (2012) [27] αναδεικνύουν ως κύρια προτερήματα του CiteSpace II, την ευελιξία που παρουσιάζει στην ανάγνωση διαφορετικών format δεδομένων, όπως αυτά των Web of Science, PubMed, arXiv και SAO/NASA, και την πληθώρα τόσο των τύπων δικτύων γνώσης που μπορεί να απεικονίσει όσο και των μεθόδων οπτικοποίησης.

Σημειώνεται όμως, πως στο συγκεκριμένο εργαλείο δεν υφίσταται προ-επεξεργασία και ποιοτικός έλεγχος μεταξύ των σταδίων άντλησης και επεξεργασίας δεδομένων, ενώ επίσης εκλείπει και η επιλογή καταχώρησης των επεξεργασμένων δεδομένων σε κάποια σχεσιακή δομή - πίνακα πριν την τελική διαδικασία της οπτικοποίησης της πληροφορίας, κάτι που θα προσέδιδε μεγαλύτερη ευελιξία στους πιθανούς τρόπους που μπορούν οι χρήστες να αξιοποιήσουν τα αποτελέσματα της βιβλιομετρικής αναζήτησης τους.

Παρόλα αυτά, αναπόφευκτο κόστος της εξασφάλισης της ευελιξίας του Bibexcel αποτελεί η σχετικά αυξημένη περιπλοκότητα στη χρήση του, όπως επισημαίνεται από τους Fredrik κ.ά. (2009) [34]. Αυτό προϋποθέτει από τον χρήστη την εξοικείωση του με τις λειτουργίες, τις συναρτήσεις και τους αλγορίθμους που αξιοποιεί το συγκεκριμένο εργαλείο. Επιπλέον, παρά το μεγάλο εύρος format αρχείων εισόδου που μπορεί να αναγνώσει, απαραίτητη κρίνεται σε κάθε περίπτωση η χειροκίνητη μορφοποίηση των αρχείων, ώστε να είναι δυνατή στη συνέχεια η δημιουργία των πινάκων συσχέτισης και επομένως, των βιβλιομετρικών δικτύων.

Για την οπτικοποίηση των δικτύων και των στατιστικών πορισμάτων της εκάστοτε επιστημονικομετρικής μελέτης, οι Sangam & Mogali (2012) [27] επισημαίνουν πως είναι απαραίτητη η χρήση τρίτων προγραμμάτων, όπως το Pajek και το SPSS, καθώς το Bibexcel εστιάζει αποκλειστικά σε μαθηματικές ρουτίνες για την διαχείριση και επεξεργασία βιβλιογραφικών μεταδεδομένων και όχι στην γραφιστική αναπαράσταση της εξαγόμενης από αυτά πληροφορίας.

2.2.10 Vantage Point

Το Vantage Point αποτελεί εργαλείο εξόρυξης κειμένου που έχει αναπτυχθεί από την Search Technology Inc. και η χρήση του βρίσκει εφαρμογή στην αναζήτηση και επεξεργασία δεδομένων από βάσεις δεδομένων επιστημονικών ευρεσιτεχνιών και δημοσιεύσεων. Στόχος του εργαλείου είναι η αναγνώριση μοτίβων και η ανάλυση των αλληλεπιδράσεων που προκύπτουν από την ανάγνωση δομημένων κειμένων.

Ιδιαίτερη έμφαση δίνεται από τους Cobo κ.ά. (2011) [26] και Sangam & Mogali (2012) [27] στο γεγονός πως το Vantage Point διαθέτει 180 φίλτρα εισαγωγής δεδομένων που δίνουν τη δυνατότητα στους χρήστες να εισάγουν δεδομένα από σχεδόν οποιαδήποτε βάση δεδομένων, με αρχεία εισόδου διαφορετικού format όπως αυτά των Microsoft Excel, Microsoft Access και XML.

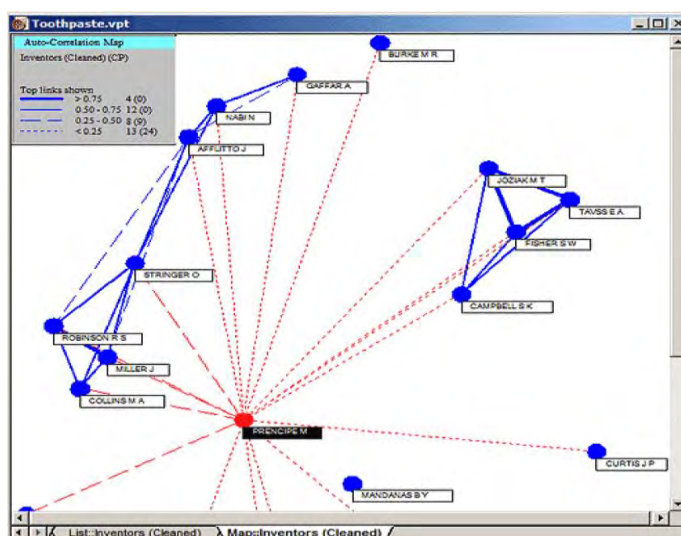
Με την ολοκλήρωση της φόρτωσης των δεδομένων στο εργαλείο, πραγματοποιείται η χαρτογράφηση (mapping) των πεδίων που εμπεριέχονται στα μεταδεδομένα στις αντίστοιχες στήλες των πινάκων όπου και θα αποθηκευτούν. Για παράδειγμα, σε μια περίπτωση φόρτωσης μεταδεδομένων επιστημονικών δημοσιεύσεων, πεδία όπως ο τίτλος, η περίληψη, οι συγγραφείς, οι οργανισμοί και οι παραπομπές του κάθε εγγράφου αντιστοιχίζονται αυτόματα στις κατάλληλες στήλες των πινάκων συσχέτισης. Επιπρόσθετα, παρέχεται και η δυνατότητα τόσο κανονικοποίησης/ομογενοποίησης στοιχείων κειμένου

όσο και η εξαγωγή σημαντικών φράσεων ή λέξεων κλειδιών που εμπεριέχονται σε στοιχεία όπως τίτλοι ή περιλήψεις δημοσιεύσεων.

Οι λειτουργίες αυτές αποδεικνύονται ιδιαίτερα χρήσιμες για την επεξεργασία των δεδομένων σε πολλαπλά στάδια. Ο Pradhan (2011) [35] παραθέτει στο άρθρο του μια κατατοπιστική περιγραφή σχετικά με τα προτερήματα που παρουσιάζει το Vantage Point στην διαχείριση και τον καθαρισμό των εισαγόμενων σε αυτό δεδομένων. Ένα από αυτά αποτελεί η δημιουργία λιστών με επιλεγμένες καταχωρήσεις οντοτήτων όπως οργανισμοί, συγγραφείς ή χώρες προέλευσης συγγραφέων. Οι λίστες αυτές μπορούν να αξιοποιηθούν για την ανάδειξη των πιο σημαντικών (βάσει βιβλιογραφικών μετρικών) καταχωρήσεων, κάτι που μειώνει το πλήθος των υπό εξέταση δεδομένων, διευκολύνοντας έτσι τη στατιστική τους ανάλυση.

Επιπλέον, μέσω της αλληλοεπικάλυψης αυτών των λιστών δημιουργούνται και οι απαραίτητοι για την εξαγωγή δικτύων γνώσης πίνακες συνύπαρξης. Προκειμένου να εξασφαλιστεί η όσο το δυνατόν υψηλότερη εγκυρότητα και ακρίβεια των εξαγόμενων δικτύων, το Vantage Point προσφέρει μεθόδους εντοπισμού και απαλοιφής διπλοτύπων. Για τον εντοπισμό πολλαπλών παραλλαγών που αντιστοιχούν στην ίδια φυσική οντότητα, εφαρμόζονται αλγόριθμοι σύγκρισης γραμματοσειρών, κάτι που αναδεικνύει την σημαντικότητα της λειτουργίας κανονικοποίησης κειμένου που αναφέρθηκε παραπάνω.

Την αποπεράτωση της διαδικασίας οργάνωσης και καθαρισμού των δεδομένων εισόδου, ακολουθεί η οπτικοποίηση των επιστημονικών δικτύων και η προβολή των στατιστικών τους στο UI του εργαλείου, όπως παρουσιάζεται και στην Εικόνα 2.11.



Εικόνα 2.11: Στιγμιότυπο από το UI του Vantage Point (πηγή: Sangam-Mogali, 2012, Mapping and Visualization Software tools a review, International Conference on Content Analysis).

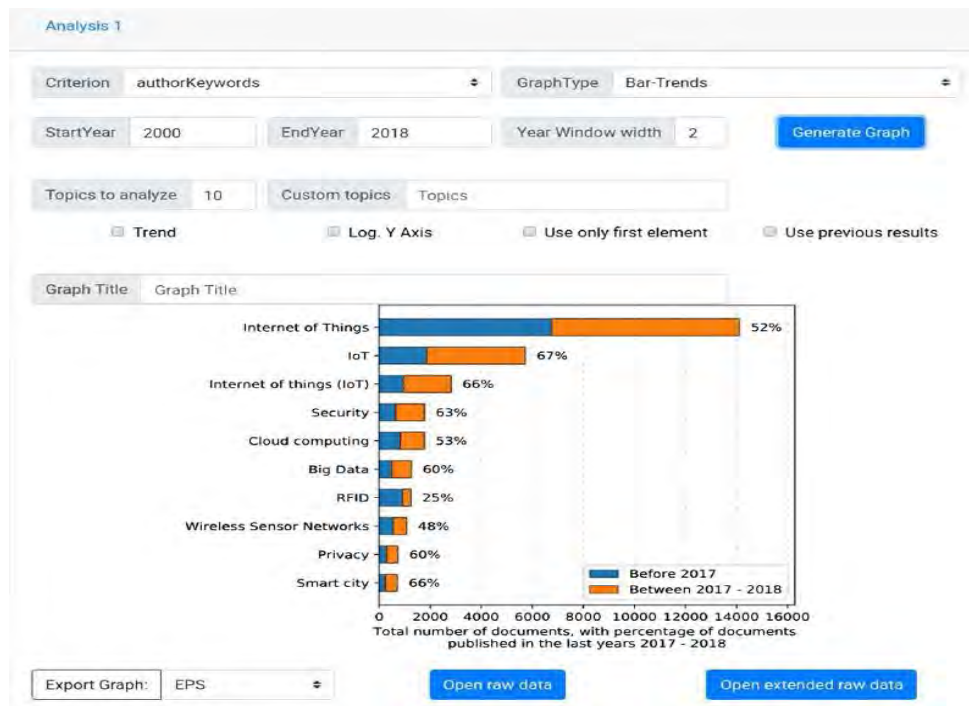
2.2.11 ScientoPy

Χρήσιμη για την αποπεράτωση της παρούσας διπλωματικής εργασίας αποτέλεσε η μελέτη της έρευνας του Ruiz-Rosero (2019) [36], μέσω της οποίας παρουσιάστηκε το ανεπτυγμένο σε Python εργαλείο βιβλιομετρικών αναλύσεων ScientoPy. Οι συγγραφείς αναλύουν την ροή διεργασιών που τελεί το ScientoPy σε τέσσερα διακριτά στάδια: άντληση, επεξεργασία, στατιστική ανάλυση και οπτικοποίηση δεδομένων.

Αρχικά πραγματοποιείται η εξαγωγή βιβλιογραφικών datasets από τις βάσεις δεδομένων των Scopus και Web of Science σε αρχεία CSV. Τα εξαγόμενα datasets εμπεριέχουν μεταδεδομένα δημοσιεύσεων όπως περιλήψεις, λέξεις – κλειδιά, ιστορικά παραπομπών της κάθε δημοσίευσης αλλά και διευθύνσεις εργασίας και χώρα διαμονής των συγγραφέων του κάθε άρθρου. Τα επιλεγμένα μεταδεδομένα συγκεντρώνονται σε έναν ενιαίο φάκελο, ώστε το ScientoPy να είναι σε θέση να τα χειριστεί κατάλληλα.

Στο στάδιο της προ-επεξεργασίας των δεδομένων, επιτελούνται λειτουργίες που αφορούν το περιεχόμενο, την οργάνωση αλλά και την ποιότητα της εισαγόμενης στο ScientoPy πληροφορίας. Συγκεκριμένα, πραγματοποιείται φιλτράρισμα των δημοσιεύσεων βάσει του τύπου τους, έτσι ώστε να παραμείνουν μόνο αυτές που ανήκουν σε συγκεκριμένους τύπους εγγράφων. Παράλληλα, πραγματοποιείται κανονικοποίηση και συγχώνευση στοιχείων κειμένου, όπως τίτλοι, πεδία, ονόματα συγγραφέων, οργανισμών και χωρών, καθώς το format τους διαφοροποιείται ανάλογα με τη βάση δεδομένων από την οποία έχουν ανακτηθεί. Τα κανονικοποιημένα μεταδεδομένα δημοσιεύσεων αξιοποιούνται, σε συνδυασμό με τα αναγνωριστικά DOI, για τον εντοπισμό εγγράφων που αντιστοιχούν στην ίδια δημοσίευση. Αφού τα διπλότυπα συγχωνευθούν σε μια ενιαία καταχώρηση, διαγράφονται. Σε περίπτωση που οι πολλαπλές εκδόσεις μιας δημοσίευσης χαρακτηρίζονται από διαφορετικό αριθμό παραπομπών, υπολογίζεται ο μέσος όρος αυτών.

Μετά τον καθαρισμό των δεδομένων, έπεται ο υπολογισμός μιας πληθώρας στατιστικών. Δείκτες όπως ο AGR, ο ADY και το PDLY υπολογίζονται για κάθε θεματικό αντικείμενο προκειμένου να αναδειχθούν οι ανερχόμενες τάσεις και τα πιο δημοφιλή επιστημονικά πεδία για ένα ορισμένο χρονικό πλαίσιο. Η δυνατότητα εισαγωγής των κριτηρίων αναζήτησης και της οπτικοποίησης των στατιστικών αποτελεσμάτων μέσω ποικίλων τύπων γραφημάτων παρέχεται μέσω του UI του εργαλείου στην Εικόνα 2.12.



Εικόνα 2.12: Παράθυρο ανάλυσης στατιστικών στο UI του ScientoPy (πηγή: Ruiz-Rosero, 2019, Software survey : ScientoPy, a scientometric tool for topics, Scientometrics).

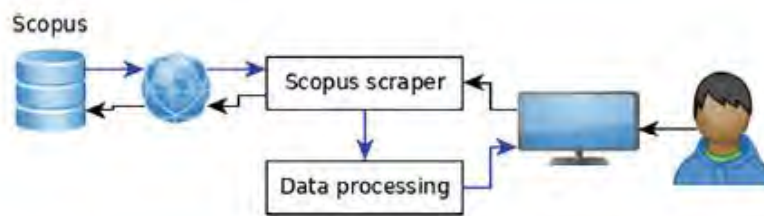
2.2.12 Pysurveillance

Οι Cestero κ.ά. (2022) [37] επισημαίνουν στο άρθρο τους την ανάγκη ανάπτυξης αποδοτικών και αυτοματοποιημένων εργαλείων, που θα αποσκοπούν στην υποστύλωση των ερευνών και αξιολογήσεων επιστημονικής βιβλιογραφίας, των οποίων η πολυπλοκότητα συνεχώς αυξάνεται λόγω της ολοένα και μεγαλύτερης συσσώρευσης δεδομένων. Στο πλαίσιο αυτό, γίνεται παρουσίαση της open-source πλατφόρμας Pysurveillance και τους πιθανούς τρόπους που μπορεί να συνεισφέρει στην εκπόνηση συστηματικών βιβλιομετρικών μελετών. Συγκεκριμένα, η Pysurveillance δρα ως διαμεσολαβητής μεταξύ της βάση δεδομένων της Scopus και των χρηστών της, κάνοντας αξιοποίηση των APIs της δεύτερης, όπως απεικονίζεται στο παρακάτω διάγραμμα ροής της Εικόνας 2.13.

Σε επίπεδο frontend προσφέρει ένα εύκολο στη χρήση dashboard μέσω του οποίου είναι δυνατή η αποστολή queries στο κέντρο διαχείρισης αιτημάτων της Scopus. Η δομή των queries μάλιστα της Pysurveillance είναι παρόμοια σκοπίμως με αυτή των queries της Scopus: αλληλουχία λέξεων κλειδιών και δυαδικών τελεστών, όπως 'AND' και 'OR'. Αυτό εξυπηρετεί τόσο την απλοποίηση της αλληλεπίδρασης χρηστών και πλατφόρμας όσο και στην αποδοτικότερη επικοινωνία των δυο APIs. Επιπρόσθετα, το UI της πλατφόρμας δίνει και τη δυνατότητα εισαγωγής δεδομένων που έχουν ανακτηθεί από τη Scopus μέσω CSV αρχείων.

Το backend της Pysurveillance αποτελείται από ένα Scopus scraper script και από ένα σύνολο μεθόδων επεξεργασίας δεδομένων. Το scraper script είναι υπεύθυνο για την παραλαβή των queries από το UI και την προώθηση τους στο API της Scopus, το οποίο τα επεξεργάζεται και επιστρέφει τα μεταδεδομένα των δημοσιεύσεων που καλύπτουν τα κριτήρια αναζήτησης. Τα δεδομένα αυτά ακολούθως επεξεργάζονται κατάλληλα με την εφαρμογή μεθόδων φιλτραρίσματος και διάταξης, ώστε τελικά να μπορούν να αναπαρασταθούν μέσω στατιστικών γραφημάτων και διαγραμμάτων.

Σύμφωνα με τους συγγραφείς, ειδοποιός διαφορά της Pysurveillance από παρόμοιου τύπου εργαλεία, αποτελεί αφενός η ευκολία και η αμεσότητα στη χρήση της πλατφόρμας μέσω του UI της και αφετέρου η δυνατότητα άμεσης άντλησης της βιβλιογραφικής πληροφορίας από την πηγή δεδομένων (την Scopus στη συγκριμένη περίπτωση) μέσω του εργαλείου. Στα μελλοντικά σχέδια ανάπτυξης της πλατφόρμας εντάσσονται και επιπρόσθετες λειτουργίες, όπως η δημιουργία καθορισμένων από τον χρήστη γραφημάτων (δυνατότητα που παρέχεται από το Bibliometrix που αναφέρθηκε παραπάνω) και η άντληση μεταδεδομένων από πολλαπλές πηγές δεδομένων και η ενοποίηση τους σε έναν ενιαίο κατάλογο.



Εικόνα 2.13: Αλληλοεπίδραση χρηστών με την πλατφόρμα της Scopus μέσω του Pysurveillance (πηγή: Cestero κ.ά., 2022, Pysurveillance : A Novel Tool, Advanced Intelligent Technologies for Industry)

2.3 Επισκόπηση εργαλείων

Για καλύτερη κατανόηση, στον [Πίνακα 2.1](#) πραγματοποιείται η επισκόπηση όλων των εργαλείων που αναφέρθηκαν σε αυτήν την υποενότητα. Στην πρώτη στήλη παρατίθεται η αναφορά του αντίστοιχου άρθρου, ενώ στη δεύτερη στήλη ο στόχος που επιδιώκεται να επιτευχθεί εντός της έρευνας. Στις επόμενες τρεις στήλες αναφέρονται αντίστοιχα το όνομα, ο τύπος και οι πηγές δεδομένων που αξιοποιήθηκαν στην εκάστοτε μελέτη περίπτωσης του εκάστοτε εργαλείου. Στις δύο τελευταίες στήλες επισημαίνονται οι δυνατότητες που παρέχονται και εκλείπουν από το λειτουργικό εύρος των εργαλείων που εξετάζονται.

Πίνακας 2.1 Επισκόπηση των βιβλιομετρικών εργαλείων που εξετάσθηκαν στην υποενότητα

Εργαλείο	Στόχος	Τύπος	Βάσεις δεδομένων	Προσφέρει	Δεν προσφέρει
Conflate	Ενοποίηση μετρικών επιστημονικής βιβλιογραφίας σε μια ενιαία πλατφόρμα	Πλατφόρμα	Scopus, Web of Science	• Εντοπισμός διπλοτύπων • Αυτόματη άντληση δεδομένων	• Code – free UI • Αυτόματη δημιουργία δομής συσχέτισης • Δυναμική διαχείριση αποτελεσμάτων
Proof of Bibliometric Indicators	Ανάπτυξη κατανεμημένου δικτύου για την καταχώρηση και την παρουσίαση μετρικών συγγραφέων	Εφαρμογή	Scopus, Web of Science, Google Scholar	• Code – free UI • Εντοπισμός διπλοτύπων • Αυτόματη άντληση δεδομένων	• Αυτόματη δημιουργία δομής συσχέτισης • Δυναμική Διαχείριση αποτελεσμάτων
Bibliometrix	Ανάλυση δεδομένων επιστημονικής βιβλιογραφίας για την εξαγωγή δικτύων γνώσης	Ψηφιακή Βιβλιοθήκη Συναρτήσεων	Scopus, Web of Science	• Code – free UI • Αυτόματη άντληση δεδομένων	• Αυτόματη δημιουργία δομής συσχέτισης • Εντοπισμός διπλοτύπων • Δυναμική διαχείριση αποτελεσμάτων

Εργαλείο	Στόχος	Τύπος	Βάσεις δεδομένων	Προσφέρει	Δεν προσφέρει
VOSviewer	Δημιουργία και αναπαράσταση βιβλιομετρικών δικτύων	Open source εργαλείο	Thompson Reuters	• Code – free UI • Δυναμική διαχείριση αποτελεσμάτων	• Εντοπισμός διπλοτύπων • Αυτόματη δημιουργία δομής συσχέτισης • Αυτόματη άντληση δεδομένων
CiteNetExplorer	Ανάλυση και απεικόνιση δίκτυα παραπομπών επιστημονικών δημοσιεύσεων	Open source εργαλείο	Web of Science	• Code – free UI • Δυναμική διαχείριση αποτελεσμάτων	• Εντοπισμός διπλοτύπων • Αυτόματη δημιουργία δομής συσχέτισης • Αυτόματη άντληση δεδομένων
SciMAT	Ανάλυση και απεικόνιση επιστημονικών δικτύων	Open source εργαλείο	Web of Science, Scopus	• Code – free UI • Εντοπισμός διπλοτύπων • Δυναμική διαχείριση αποτελεσμάτων	• Αυτόματη δημιουργία δομής συσχέτισης • Αυτόματη άντληση δεδομένων

Άρθρο	Στόχος	Τύπος	Βάσεις δεδομένων	Προσφέρει	Δεν προσφέρει
Sci2 Tool	Ανάλυση, επεξεργασία και οπτικοποίηση δεδομένων επιστημονικής βιβλιογραφίας	Open source εργαλείο	Web of Science	• Code – free UI • Εντοπισμός διπλοτύπων • Δυναμική διαχείριση αποτελεσμάτων	• Αυτόματη δημιουργία δομής συσχέτισης • Αυτόματη άντληση δεδομένων
CiteSpace II	Εντοπισμός, ανάλυση και οπτικοποίηση μοτίβων και τάσεων στην επιστημονική βιβλιογραφία	Open source εργαλείο	Web of Science, PubMed	• Code – free UI	• Εντοπισμός διπλοτύπων • Δυναμική διαχείριση αποτελεσμάτων • Αυτόματη δημιουργία δομής συσχέτισης • Αυτόματη άντληση δεδομένων
Bibexcel	Διαχείριση και επεξεργασία βιβλιομετρικών δεδομένων για την εξαγωγή δικτύων γνώσης	Open source εργαλείο	Scopus, Web of Science και Procite	• Code – free UI • Εντοπισμός διπλοτύπων • Δυναμική διαχείριση αποτελεσμάτων	• Αυτόματη δημιουργία δομής συσχέτισης • Αυτόματη άντληση δεδομένων

Άρθρο	Στόχος	Τύπος	Όνομα	Βάσεις δεδομένων	Προσφέρει	Δεν προσφέρει
Vantage Point	Εξόρυξη στοιχείων κειμένου για την δημιουργία δικτύων γνώσης	Open source εργαλείο	Vantage Point	Web of Science	• Code – free UI • Εντοπισμός διπλοτύπων • Δυναμική διαχείριση αποτελεσμάτων	• Αυτόματη δημιουργία δομής συσχέτισης • Αυτόματη άντληση δεδομένων
ScientoPy	Ανάλυση βιβλιομετρικών δεδομένων και απεικόνιση τάσεων σε επιστημονικούς τομείς	Open source εργαλείο	ScientoPy	Scopus, Web of Science	• Code – free UI • Εντοπισμός διπλοτύπων	• Δυναμική διαχείριση αποτελεσμάτων • Αυτόματη δημιουργία δομής συσχέτισης • Αυτόματη άντληση δεδομένων
Pysurveillance	Οπτική ανάλυση δεδομένων σχετικά με κάποιον επιστημονικό τομέα ή αντικείμενο	Open source εργαλείο	Pysurveillance	Scopus	• Code – free UI • Εντοπισμός διπλοτύπων • Αυτόματη άντληση δεδομένων	• Αυτόματη δημιουργία δομής συσχέτισης • Δυναμική διαχείριση αποτελεσμάτων

2.4 Συμπεράσματα από την ανασκόπηση σχετικών ερευνών και εργαλείων

Από την περιγραφή των εργαλείων ανάλυσης δεδομένων επιστημονικής βιβλιογραφίας στην Ενότητα 2.2 και την επισκόπηση των δυνατοτήτων αλλά και των περιορισμών τους στην Ενότητα 2.3, εξάγονται συμπεράσματα σχετικά με το ποιες από τις επιμέρους διαστάσεις του θέματος της διπλωματικής εργασίας που αναφέρθηκαν στην Ενότητα 2.4, είναι αυτές στις οποίες παρέχεται μεγαλύτερη ή μικρότερη κάλυψη.

Στο Διάγραμμα 2.1 διαφαίνονται οι λειτουργίες που λήφθηκαν υπόψιν κατά την αξιολόγηση ερευνών στην υπάρχουσα βιβλιογραφία και τα αντίστοιχα ποσοστά κάλυψης από τα εργαλεία που έχουν αναπτυχθεί και βρίσκουν εφαρμογή στην εκπόνηση στατιστικών αναλύσεων επιστημονικής συγγραφής. Όπως παρατηρείται, ο σχεδιασμός ενός code – free UI απαντάται στην πλειονότητα των περιπτώσεων που διερευνήθηκαν, ούτως ώστε η αξιοποίηση τους από τους χρήστες να είναι όσο το δυνατόν πιο κατανοητή και απλή, χωρίς την ανάγκη γνώσης προγραμματισμού.

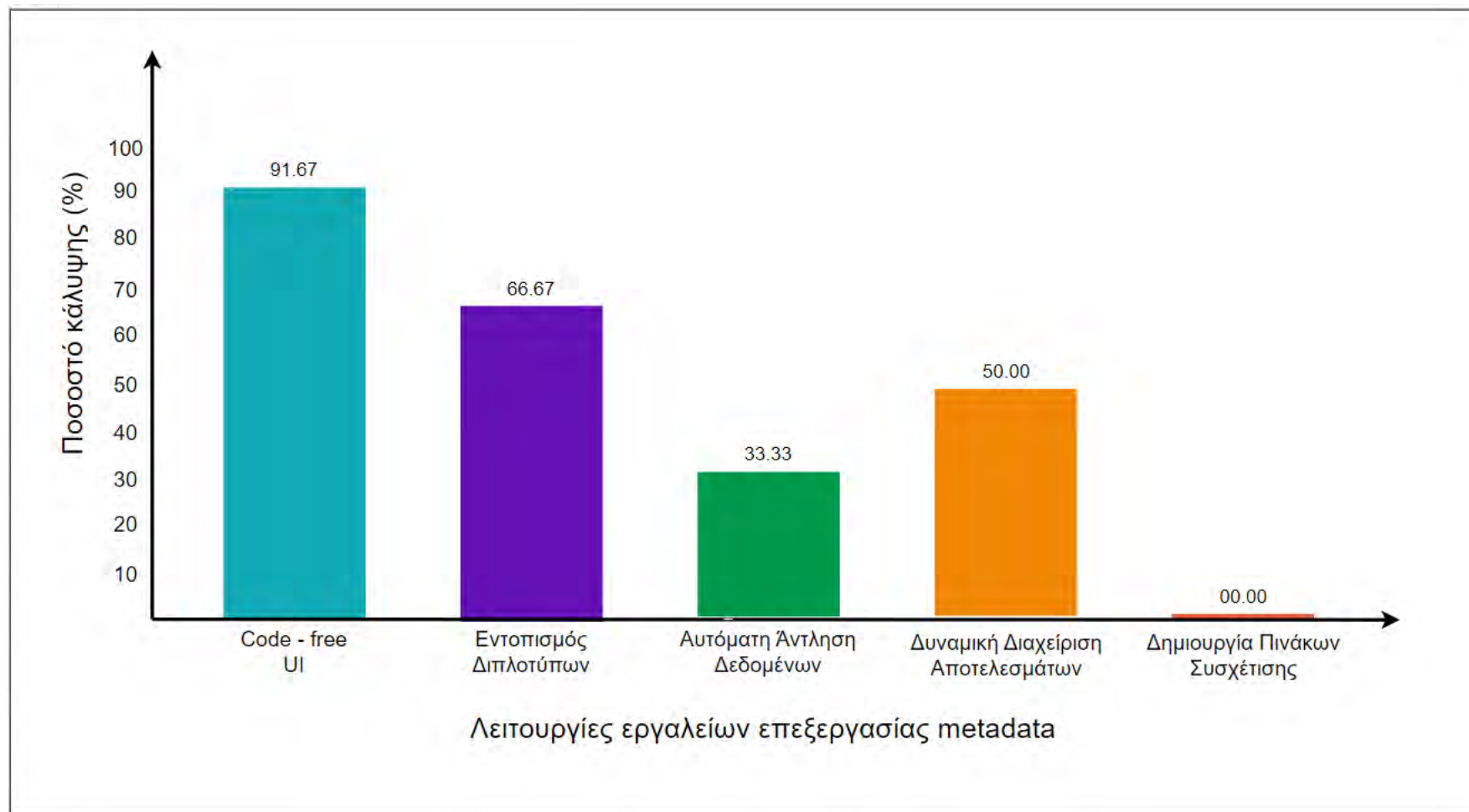
Κατά προσέγγιση, το 67% των εργαλείων που αναφέρθηκαν παρέχουν δυνατότητα εντοπισμού διπλοτύπων, με το εύρος των τεχνικών να ποικίλει από αξιοποίηση επίσημων αναγνωριστικών (π.χ. DOI, ORCID ID) που παρέχουν αυξημένη αξιοπιστία, μέχρι αλγόριθμους σύγκρισης στοιχείων κειμένου, οι οποίοι αυξάνουν σημαντικά το πλήθος των διαθέσιμων δεδομένων προς επεξεργασία, καθώς καθιστούν μη αναγκαίο τα ψηφιακά έγγραφα ή τα profiles ερευνητών να φέρουν κάποιο μοναδικό χαρακτηριστικό.

Ένα συγκρίσιμο ποσοστό κάλυψης του 50% παρατηρείται και στην ενσωμάτωση της δυνατότητας δυναμικής διαχείρισης των αποτελεσμάτων στο λειτουργικό εύρος των λογισμικών. Η επέκταση των πιθανών τρόπων αξιοποίησης των επεξεργασμένων δεδομένων, πέρα από την εξαγωγή στατιστικών πορισμάτων, αφορά σε αυτές τις περιπτώσεις εργαλείων και το φιλτράρισμα ή/και την αναζήτηση συγκεκριμένων καταγραφών, την επιλογή των αντιγράφων προς διαγραφή ή συγχώνευση και τη δημιουργία αρχείων εξόδου, που να επιτρέπουν την αξιοποίηση των τελικών δεδομένων και από άλλα αντίστοιχου τύπου εργαλεία.

Ένα μικρότερο ποσοστό κάλυψης αντιστοιχεί στην αυτοματοποίηση της διαδικασίας άντλησης δεδομένων. Στις περισσότερες έρευνες η εισαγωγή των πρωτογενών δεδομένων στα εργαλεία γινόταν χειροκίνητα μέσω ογκωδών και δύσκολων στην επεξεργασία αρχείων, ενώ σε λιγότερες η διαδικασία αυτοματοποιήθηκε μέσω της ανάπτυξης κώδικα και της αξιοποίησης του API της εκάστοτε πηγής δεδομένων επιστημονικής βιβλιογραφίας.

Τέλος, σε καμία περίπτωση εκ των εργαλείων που εξετάσθηκαν δεν παρατηρείται λειτουργία αυτόματης οργάνωσης των metadata επιστημονικής βιβλιογραφίας σε μια ενιαία σχεσιακή δομή πίνακα, που να επιτρέπει την στατιστική τους ανάλυση. Η έλλειψη αυτής της λειτουργίας έχει ως αποτέλεσμα την χειροκίνητη επεξεργασία και οργάνωση σε κατάλληλη μορφή των metadata από τους ίδιους τους χρήστες. Αυτό σημαίνει ότι την ανάκτηση των δεδομένων, θα πρέπει να ακολουθήσει η διερεύνηση για την κάθε δημοσίευση ξεχωριστά με το ποια ομάδα ερευνητών σχετίζεται η συγγραφή της. Αντίστοιχα, θα πρέπει να διερευνηθούν και οι διασυνδέσεις του κάθε συγγραφέα με τους οργανισμούς στους οποίους δραστηριοποιείται στον χρόνο συγγραφής της εκάστοτε δημοσίευσης. Γίνεται αντιληπτό πως η απουσία αυτοματοποίησης στην διεξαγωγή της παραπάνω διαδικασίας, για datasets που περιέχουν metadata εκατοντάδων χιλιάδων ή και εκατομμυρίων δημοσιεύσεων, την καθιστά ιδιαίτερα χρονοβόρα, δύσκολη και επιρρεπή σε σφάλματα.

Το αντικείμενο της παρούσας διπλωματικής εργασίας στρέφεται προς την ανάπτυξη μιας διαδικτυακής πλατφόρμας που θα προσφέρει αφενός την δυνατότητα αναζήτησης και άντλησης metadata επιστημονικών δημοσιεύσεων και αφετέρου την αυτοματοποιημένη συσχέτιση των οντοτήτων ακαδημαϊκής συγγραφής εντός δομής πίνακα. Θέτεται επίσης και ως στόχος, η ανάπτυξη μεθόδων καθαρισμού των δεδομένων από διπλότυπα, προκειμένου να είναι εξασφαλισμένη η ποιότητα των metadata προς επεξεργασία. Με αυτόν τον τρόπο επιτυγχάνεται δραστική μείωση του χρόνου και της δυσκολίας της οργάνωσης δεδομένων σε κατάλληλες δομές, χωρίς την ύπαρξη εκπτώσεων στην αξιοπιστία και την ποιότητα της πληροφορίας που εμπεριέχεται σε αυτά τα σύνολα δεδομένων.



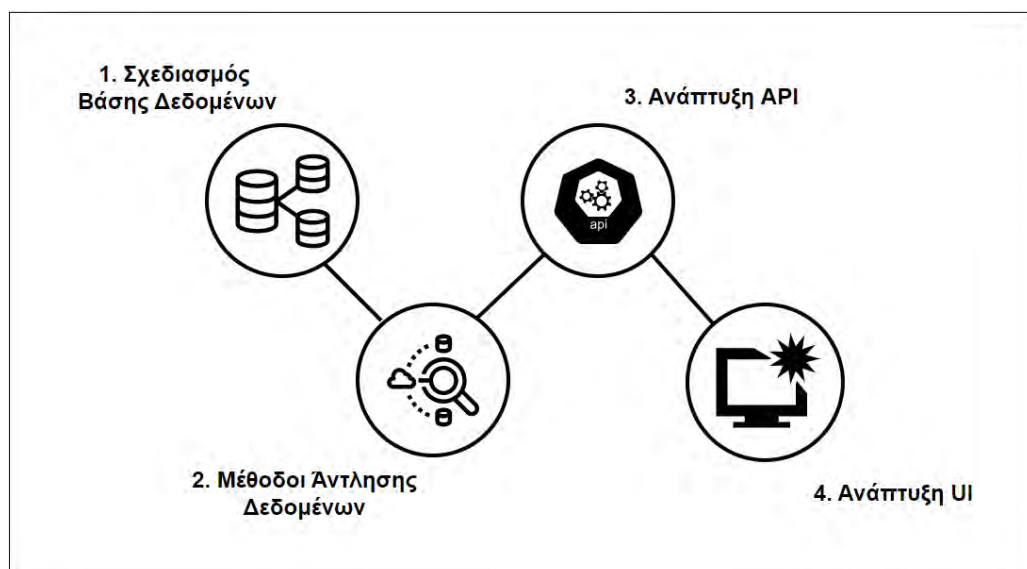
Διάγραμμα 2.1 Κάλυψη των λειτουργιών από εργαλεία επεξεργασίας δεδομένων επιστημονικής βιβλιογραφίας (πηγή: ιδία χρήση).

Κεφάλαιο 3 ΑΝΑΠΤΥΞΗ ΜΕΘΟΔΟΛΟΓΙΑΣ

3.1 Ανάλυση μεθοδολογίας σε επιμέρους στάδια

Στην Εικόνα 3.1 παρουσιάζονται τα βασικά στάδια της μεθοδολογίας που ακολουθήθηκε στην παρούσα διπλωματική εργασία, τα οποία είναι:

1. Σχεδιασμός βάσης δεδομένων, που αφορά τον καθορισμό των οντοτήτων – πινάκων, τα χαρακτηριστικά τους, τις μεταξύ τους συσχετίσεις καθώς και τον καθορισμό των απαραίτητων περιορισμών (περιορισμοί πρωτευόντων κλειδιών, ξένα κλειδιά, μοναδικών χαρακτηριστικών κ.ά.).
2. Επιλογή πηγών δεδομένων, κατά την οποία εξετάζονται κοινόχρηστες πλατφόρμες και βάσεις δεδομένων που προσφέρουν πρόσβαση σε metadata επιστημονικών δημοσιεύσεων μέσω των APIs που διαθέτουν.
3. Ανάπτυξη API για την τέλεση διεργασιών όπως:
 - Η ανάκτηση metadata επιστημονικής βιβλιογραφίας από τις πηγές δεδομένων που έχουν επιλεγεί στο Στάδιο 2.
 - Η κατάλληλη προ-επεξεργασία των άμεσα αντλούμενων δεδομένων.
 - Η καταχώρηση τους στην βάση δεδομένων που αναπτύχθηκε στο Στάδιο 1.
 - Ο εντοπισμός πολλαπλών καταχωρήσεων που αντιστοιχούν στην ίδια φυσική οντότητα.
4. Ανάπτυξη UI που θα αποστέλλει τα αιτήματα των χρηστών στο API του Σταδίου 4 και θα δίνει τη δυνατότητα προεπισκόπησης αλλά και επεξεργασίας των δεδομένων που ανακτώνται σε κάθε περίπτωση.



Εικόνα 14: Διάγραμμα ροής διεργασιών που ακολουθήθηκαν (πηγή: ιδία χρήση)

3.2 Ανάπτυξη βάσης δεδομένων

3.2.1 Επιλογή RDBMS

Ως RDBMS που θα φιλοξενήσει το σχήμα της βάσης δεδομένων που αναπτύσσεται, επιλέχθηκε το client/server σύστημα MySQL Community Server.

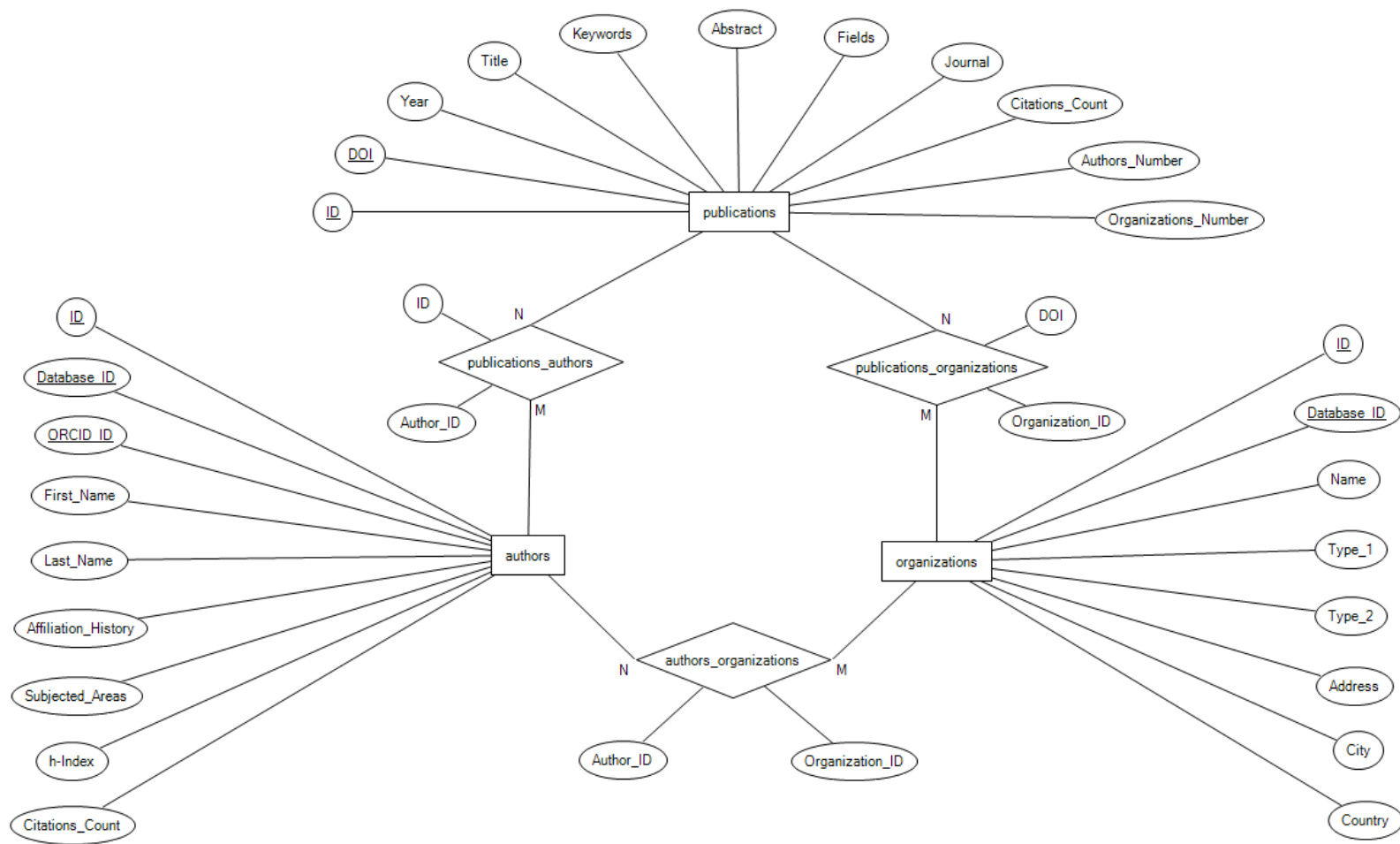
Το συγκεκριμένο σύστημα προσφέρει:

- MySQL Command Line, για την εκτέλεση εντολών σχετικά με τη δημιουργία νέων χρηστών και βάσεων δεδομένων μέσω γραμμής εντολών.
- MySQL Shell, που αποτελεί μια γραφιστική επιφάνεια εργασιών και επιτρέπει ενέργειες όπως σύνταξη και εκτέλεση SQL queries, δημιουργία νέων σχημάτων βάσεων δεδομένων και διαχείριση των λογαριασμών των χρηστών.
- MySQL Server, ώστε να είναι δυνατή η εγκαθίδρυση HTTP πρωτοκόλλου επικοινωνίας με APIs άλλων λογισμικών.



3.2.2 Περιγραφή της δομής της βάσης δεδομένων

Στην Εικόνα 3.2 παρουσιάζεται το διάγραμμα συσχέτισης των βασικών οντοτήτων μέσω των οποίων πραγματοποιείται η οργάνωση και η διασύνδεση των βιβλιομετρικών δεδομένων που αντλούνται. Το διάγραμμα δημιουργήθηκε από το διαδικτυακό εργαλείο μοντελοποίησης βάσεων δεδομένων ER2SQL, που αναπτύχθηκε από τους Παπαδοπούλου & Βασιλακόπουλο (2017) [38].



Εικόνα 3.2: Σχεδιασμός διαγράμματος συσχέτισης οντοτήτων μέσω ER2SQL (πηγή: ίδια χρήση).

Οι τρεις βασικές οντότητες και τα αντίστοιχα χαρακτηριστικά τους που παρουσιάζονται στο παραπάνω διάγραμμα είναι:

- Δημοσιεύσεις, με χαρακτηριστικά:
 - ο Μοναδικό αναγνωριστικό ID (πρωτεύον κλειδί)
 - ο DOI - μοναδικό χαρακτηριστικό
 - ο Έτος συγγραφής
 - ο Περιοδικό (ή εκδοτικός οργανισμός/συνέδριο)
 - ο Τίτλος
 - ο Λέξεις – κλειδιά
 - ο Περίληψη
 - ο Πεδίο
 - ο Αριθμός Αναφορών
 - ο Αριθμός Συγγραφέων
 - ο Αριθμός Οργανισμών
- Συγγραφείς, με χαρακτηριστικά:
 - ο Μοναδικό αναγνωριστικό ID (πρωτεύον κλειδί)
 - ο Αναγνωριστικό από την βάση δεδομένων από την οποία αντλήθηκαν τα metadata του profile του - μοναδικό χαρακτηριστικό.
 - ο ORCID ID
 - ο Όνομα
 - ο Επώνυμο
 - ο Πεδία Μελέτης
 - ο Διασυνδεδεμένοι οργανισμοί
 - ο h-Index
 - ο Αριθμός Αναφορών
- Οργανισμοί, με χαρακτηριστικά:
 - ο Μοναδικό αναγνωριστικό ID (πρωτεύον κλειδί)
 - ο Αναγνωριστικό από την βάση δεδομένων από την οποία αντλήθηκαν τα metadata του profile του – μοναδικό χαρακτηριστικό.
 - ο Όνομα
 - ο Τύπος 1
 - ο Τύπος 2
 - ο Διεύθυνση
 - ο Πόλη
 - ο Χώρα

Πέραν αυτών των πινάκων που περιέχουν την βιβλιομετρική πληροφορία, απαραίτητη είναι και η συμπερίληψη των πινάκων συσχέτισης μεταξύ των οντοτήτων στο σχήμα της βάσης δεδομένων. Συγκεκριμένα, μέσω αυτών των πινάκων πραγματοποιείται συσχέτιση:

- Δημοσιεύσεων – Συγγραφέων, με ξένα κλειδιά τα μοναδικά αναγνωστικά (ID) των οντοτήτων των Δημοσιεύσεων και Συγγραφέων αντίστοιχα.
- Δημοσιεύσεων – Οργανισμών, με ξένα κλειδιά τα μοναδικά αναγνωριστικά (ID) των οντοτήτων των Δημοσιεύσεων και Οργανισμών αντίστοιχα.
- Συγγραφέων – Οργανισμών, με ξένα κλειδιά τα μοναδικά αναγνωριστικά (ID) των οντοτήτων των Συγγραφέων και Οργανισμών αντίστοιχα.

Κάθε καταχώρηση δημοσίευσης χαρακτηρίζεται από έναν αριθμό συγγραφέων και από έναν αριθμό οργανισμών. Τα συγκεκριμένα χαρακτηριστικά μπορούν να αξιοποιηθούν για τον υπολογισμό του ποσοστού των επιτυχών καταχωρήσεων metadata συγγραφέων και οργανισμών στη βάση δεδομένων. Για αυτόν τον λόγο, ως δείκτης ακεραιότητας δεδομένων για τις καταγραφές συγγραφέων ορίζεται το κλάσμα,

$$\frac{\Sigma P_{auth}}{\Sigma P_{rel}}$$

ενώ για τις καταγραφές οργανισμών,

$$\frac{\Sigma P_{org}}{\Sigma P_{rel}}$$

Οι δύο αριθμητές των παραπάνω κλασμάτων αντιστοιχούν στο άθροισμα του αριθμού των συγγραφέων και οργανισμών αντίστοιχα για όλες τις δημοσίευσης. Οι παρονομαστές αποτελούν το άθροισμα των καταγραφών (σειρών) του αντίστοιχου πίνακα συσχέτισης (Δημοσιεύσεων – Συγγραφέων για τους συγγραφείς και Δημοσιεύσεων – Οργανισμών για τους οργανισμούς).

Κάθε δημοσίευση διασυνδέεται με έναν αριθμό συγγραφέων, ο οποίος καταχωρείται στο πεδίο ‘Αριθμός Συγγραφέων’ του πίνακα Δημοσιεύσεων. Στην περίπτωση που τα metadata όλων των συγγραφέων για κάποια δημοσίευση έχουν καταχωρηθεί επιτυχώς στη βάση δεδομένων, το αναγνωριστικό της συγκεκριμένης δημοσίευσης θα πρέπει να συναντάται τόσες φορές στον πίνακα Δημοσιεύσεων – Συγγραφέων όσο είναι και ο αριθμός των συγγραφέων της στο αντίστοιχο πεδίο του πίνακα Δημοσιεύσεων. Μέσω του τελεστή του αθροίσματος, αυτός ο έλεγχος πραγματοποιείται συνολικά για όλες τις δημοσιεύσεις.

Αντίστοιχη είναι και η λογική για τον υπολογισμό του δείκτη ακεραιότητας metadata οργανισμών.

Ακόμη δύο χαρακτηριστικά που καθορίζονται από την επεξεργασία των αντλούμενων δεδομένων είναι οι Τύποι 1 και 2, βάσει των οποίων πραγματοποιείται η κατηγοριοποίηση των οργανισμών σε δύο επίπεδα. Οι επιμέρους κατηγορίες οργανισμών στον Τύπο 1 είναι:

- Ακαδημαϊκός οργανισμός
- Κυβερνητικός οργανισμός
- Μη κερδοσκοπικός οργανισμός
- Σύλλογος/Οργάνωση/Ένωση
- Επιχείρηση

Η κατηγοριοποίηση Τύπου 2 αφορά την ομαδοποίηση των ακαδημαϊκών οργανισμών, με τις επιμέρους κατηγορίες να είναι:

- Πανεπιστήμιο / Κολλέγιο
- Σχολή
- Ακαδημία
- Ερευνητικό Κέντρο – Ινστιτούτο

Η εισαγωγή των δύο αυτών χαρακτηριστικών αποσκοπεί στον εμπλουτισμό της βιβλιομετρικής πληροφορίας, καθώς καθιστά δυνατό τον υπολογισμό θεσμικής εγγύτητας μεταξύ των οργανισμών που αλληλοεπιδρούν για την παραγωγή επιστημονικού έργου. Περισσότερες λεπτομέρειες σχετικά με τα κριτήρια και τις μεθόδους κατηγοριοποίησης των οργανισμών, θα δοθούν σε επομένη παράγραφο.

Σε αυτό το σημείο αξίζει να σημειωθεί πως μία χρήσιμη λειτουργία που προσφέρει το ER2SQL είναι η αυτόματη εξαγωγή SQL queries από το διάγραμμα πινάκων και συσχετίσεων που έχει σχεδιαστεί, κάτι που μειώνει δραστικά την διάρκεια της διαδικασίας της δημιουργίας πινάκων κατά τον σχηματισμό του σχήματος βάσης.

3.3 Άντληση metadata επιστημονικής βιβλιογραφίας

3.3.1 Διερεύνηση διαθέσιμων πηγών δεδομένων

Στους Πίνακες 3.1, 3.2 και 3.3 παρατίθενται όλες οι βάσεις δεδομένων επιστημονικής βιβλιογραφίας για τις οποίες εξετάστηκαν τα metadata δημοσιεύσεων, συγγραφέων και οργανισμών (στα οποία δίδεται πρόσβαση μέσω των αντίστοιχων APIs) καθώς και οι όροι χρήσης των συγκεκριμένων APIs που θέτονται από την κάθε πλατφόρμα.

	DOI	Έτος	Τίτλος	Περίληψη	Λέξεις- κλειδιά	Πεδία	Αναφορές
Scopus	✓	✓	✓	✓	✓	✓	✓
WoS	✓	✓	✓	✓	✓	✓	✓
Lens	✓	✓	✓	✓	✓	✓	✓
Dimensions	✓	✓	✓	✗	✗	✓	✓
SerpAPI (GS)	✗	✓	✓	✗	✗	✗	✓

Πίνακας 3.1 Διαθέσιμα metadata δημοσιεύσεων από πλατφόρμες επιστημονικής βιβλιογραφίας

	Όνομα	Επώνυμο	ORCID ID	Πεδία	Διασυνδέσεις	h-Index	Αναφορές
Scopus	✓	✓	✓	✓	✓	✓	✓
WoS	✓	✓	✓	✓	✓	✓	✓
Lens	✓	✓	✓	✓	✓	✗	✓
Dimensions	✓	✓	✓	✗	✓	✗	✓
SerpAPI (GS)	✓	✓	✗	✓	✗	✓	✗

Πίνακας 3.2 Διαθέσιμα metadata συγγραφέων από πλατφόρμες επιστημονικής βιβλιογραφίας

	Όνομα	Διεύθυνση	Πόλη	Χώρα
Scopus	✓	✓	✓	✓
WoS	✓	✓	✓	✓
Lens	✓	✓	✓	✓
Dimensions	✓	✗	✓	✓
SerpAPI (GS)	✗	✗	✗	✗

Πίνακας 3.3 Διαθέσιμα metadata οργανισμών από πλατφόρμες επιστημονικής βιβλιογραφίας

Παρατηρείται πως το μεγαλύτερο πλήθος metadata προσφέρεται από τις πλατφόρμες των Scopus και Web of Science, οι οποίες μάλιστα παρέχουν πλήρη κάλυψη εκείνων των χαρακτηριστικών που καθορίστηκαν ως απαραίτητα για τους σκοπούς της διπλωματικής εργασίας, ενώ τα APIs των Lens, Dimensions και Google Scholar δίνουν πρόσβαση σε μικρότερο πλήθος δεδομένων. Πρέπει να σημειωθεί πως στην περίπτωση της Google

Scholar, πρόσβαση στους καταλόγους δεδομένων της επιτυγχάνεται μέσω του third-party API SerpAPI, καθώς η ίδια η πλατφόρμα δεν παρέχει κάποιο επίσημη υπηρεσία άντλησης δεδομένων μέσω λογισμικού.

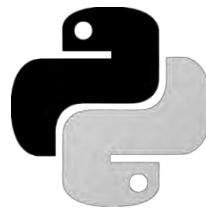
3.3.2 Επιλογή πηγής δεδομένων

Για την αποπεράτωση της διπλωματικής εργασίας, εγκρίθηκε πρόσβαση στους καταλόγους δεδομένων της Scopus μέσω απόκτησης API Key. Το API Key παρέχει αυθεντικοποίηση για τη χρήση του API της πλατφόρμας, επιτρέποντας την εξαγωγή και αξιοποίηση metadata μέσω λογισμικού. Η Scopus προσφέρει ικανοποιητική κάλυψη για τις ανάγκες της έρευνας σε πρωτογενή δεδομένα, οπότε αποφασίστηκε να χρησιμοποιηθεί η βάση δεδομένων της Scopus για την ανάκτηση metadata δημοσιεύσεων ακαδημαϊκής βιβλιογραφίας.

3.3.3 Εργαλεία που χρησιμοποιήθηκαν

Για την εξαγωγή metadata από την συλλογή δεδομένων της Scopus αναπτύχθηκε κώδικας Python στον οποίο ενσωματώθηκαν τα πακέτα:

- requests, για την πραγματοποίηση HTTP μεθόδων ανάμεσα στο Python script που αναπτύσσεται και το API της Scopus.
- json, για την μορφοποίηση των ανακτώμενων δεδομένων σε JSON format για τον αποτελεσματικότερο χειρισμό τους.
- pybliometrics, το οποίο προσφέρει εξειδικευμένες κλάσεις και μεθόδους για την ανάκτηση, την οργάνωση και τον χειρισμό δεδομένων που ανακτώνται από την Scopus.



3.3.4 Άντληση metadata – Στάδιο 1

Η διαδικασία της άντλησης δεδομένων διακρίνεται σε δύο στάδια. Στο πρώτο στάδιο, τα κριτήρια αναζήτησης που καλείται να εισάγει ο χρήστης ενσωματώνονται σε HTTP query το οποίο αποστέλλεται στο API της Scopus, με σκοπό την ανάκτηση του αναγνωριστικού DOI κάθε δημοσίευσης που είναι συναφής με τις αρχικές παραμέτρους αναζήτησης. Στο δεύτερο στάδιο, το αναγνωριστικό DOI αξιοποιείται κατάλληλα για την άντληση και την οργάνωση των metadata.

Συγκεκριμένα, για την υλοποίηση του πρώτου σταδίου αναπτύχθηκε η μέθοδος getDois που παρουσιάζεται στις δύο επόμενες εικόνες. Στην Εικόνα 3.3 προβάλλονται οι παράμετροι εισόδου και η αρχικοποίηση των μεταβλητών που χρησιμοποιούνται από την μέθοδο. Οι παράμετροι εισόδου αποτελούν τα κριτήρια αναζήτησης που καλείται να εισάγει ο χρήστης

όπως λέξεις – κλειδιά, τελεστές Boolean που καθορίζουν τις συσχετίσεις των λέξεων – κλειδιών μεταξύ τους, έτος έκδοσης και επιστημονικά πεδία που είναι επιθυμητό να συμπεριλαμβάνονται στα αντίστοιχα πεδία των ανακτώμενων δημοσιεύσεων. Επιπρόσθετα, απαραίτητη για την επιτυχή εκτέλεση της μεθόδου είναι και η εισαγωγή του API Key εκ μέρους της χρήστη, ώστε η πρόσβαση του προγράμματος στο περιεχόμενο της Scopus να αυθεντικοποιηθεί.

```
def getDois(keywords, yearPublished, fields, booleans, apiKey):  
  
    # DOIs list  
    dois = []  
  
    # query parameters  
    count = '&count=25'  
    terms = f'({buildKeywordsQuery(keywords, booleans)})'  
    scope = 'TITLE-ABS-KEY'  
    view = '&view=standard'  
    sort = '&sort=citedby_count'  
    date = '&date=' + str(yearPublished)  
    scopusAPIKey = f'&apiKey={apiKey}'  
    scopusBaseUrl = 'http://api.elsevier.com/content/search/scopus?'
```

Εικόνα 3.3: Ορισμός μεθόδου getDois και αρχικοποίηση μεταβλητών (πηγή: ιδία χρήση).

Στις μεταβλητές που αρχικοποιούνται με την έναρξη της εκτέλεσης της μεθόδου περιλαμβάνεται η λίστα στην οποία θα αποθηκευτούν τα DOIs των δημοσιεύσεων και οι παράμετροι που θα ενσωματωθούν στο query. Στις παραμέτρους αυτές συγκαταλέγεται η μεταβλητή:

- count, που καθορίζει τον αριθμό των αποτελεσμάτων που επιστρέφονται σε κάθε επανάληψη
- terms, που περιλαμβάνει τις λέξεις – κλειδιά και τους τελεστές Boolean.
- scope, που καθορίζει τις δομές στις οποίες θα αναζητηθούν οι λέξεις – κλειδιά (τίτλος, περίληψη και λέξεις – κλειδιά στη συγκεκριμένη περίπτωση).
- sort, που είναι το κριτήριο κατάταξης των επιστρεφόμενων αποτελεσμάτων (βάσει αριθμού παραπομπών εν προκειμένω).
- date, που αντιστοιχεί στο έτος έκδοσης.
- scopusAPIKey και scopusBaseUrl, που αποτελούν τα Scopus API Key και online σύνδεσμο αντίστοιχα.

Μετά την αρχικοποίηση των μεταβλητών, η εκτέλεση του προγράμματος διατρέχει κάθε επιστημονικό πεδίο που περιλαμβάνεται στην αντίστοιχη παράμετρο εισόδου, όπως δείχνεται στην Εικόνα 3.4. Η μεταβλητή errorCounter χρησιμοποιείται για την απαρίθμηση

των σφαλμάτων που είναι δυνατόν να προκύψουν κατά τη διάρκεια μιας επανάληψης. Σε περίπτωση που η μεταβλητή αποκτήσει τιμή μεγαλύτερη από ένα ανώτατο όριο, η αναζήτηση δημοσιεύσεων συνεχίζεται βάσει του επόμενου επιστημονικού πεδίου αντί να διακοπεί εντελώς. Η μεταβλητή `startIndex` αντιστοιχεί στη δημοσίευση από την οποία ξεκινάει η αναζήτηση σε κάθε νέα επανάληψη.

```
for field in tqdm(fields):

    errorCounter = 0
    startIndex = 0
    while True:
        start = f"&start={startIndex}"
        currentField = f"&subj={field}"

        # building GET query
        query = 'query=' + scope + terms + date + start + \
            count + sort + currentField + scopusAPIKey + view
        url = scopusBaseUrl + query

        req = get(url)
        # if request is successful, get DOIs
        if req.status_code == 200:
            content = json.loads(req.content)['search-results']
            totalResults = int(content['opensearch:totalResults'])
            startIndex = int(content['opensearch:startIndex'])
            entries = content['entry']
            # else print the error cause
        else:
            Error = json.loads(req.content)['service-error']['status']
            print(req.status_code, Error['statusText'])
            errorCounter += 1

        for entry in entries:
            try:
                TempDOI = entry['prism:doi']
                dois.append(str(TempDOI))
            except:
                pass

        remainingData = totalResults - startIndex - len(entries)

        # if there are any records remained, update startIndex and start the next loop
        if ((remainingData > 0) & (errorCounter < 10)):
            startIndex += 25
        # else exit the loop and continue with the next subject
        else:
            break

    return dois
```

Εικόνα 15: Σχηματισμός HTTP query για την ανάκτηση αναγνωριστικών DOI (πηγή: ιδία χρήση).

Εντός της κύριας ακολουθίας των `for` επαναλήψεων, εκτυλίσσεται και μια δεύτερη ακολουθία `while` επαναλήψεων. Αυτό είναι απαραίτητο γιατί σε πολλές περιπτώσεις τα επιστρεφόμενα αποτελέσματα για ένα πεδίο να ξεπερνούν το μέγιστο όριο πλήθους αποτελεσμάτων (25 σύμφωνα με τα δικαιώματα πρόσβασης με το API Key που χρησιμοποιήθηκε) ανά αίτημα στο API της Scopus. Επομένως, το συγκεκριμένο πεδίο πρέπει να προσπελαστεί πολλαπλές φορές ώστε να ανακτηθούν όλα τα αντικείμενα που αντιστοιχούν σε αυτό. Η μεταβλητή `startIndex` αποθηκεύει την θέση του τελευταίου άρθρου στους καταλόγους της Scopus, ώστε η αναζήτηση να συνεχίσει από το επόμενο στην επανάληψη που ακολουθεί.

Σε κάθε επανάληψη, επαναδομείται το query με το νέο startIndex και πεδίο (σε περίπτωση που το τελευταίο έχει αλλάξει) και στέλνεται στο API της Scopus μέσω HTTP αιτήματος που πραγματοποιείται από τη μέθοδο get του πακέτου requests. Σε περίπτωση που το αίτημα είναι επιτυχές (200 κωδικός κατάστασης), τα αποτελέσματα που επεστράφησαν οργανώνονται σε JSON δομή από την οποία τελικά εξάγονται ο νέος startIndex, τα συνολικά αποτελέσματα της αναζήτησης και το αναγνωριστικό DOI του κάθε paper. Αν το αίτημα είναι ανεπιτυχές, τυπώνεται η αιτία σφάλματος στη γραμμή εντολών και ο καταμετρητής σφαλμάτων αυξάνεται κατά μία μονάδα.

Σε κάθε επανάληψη υπολογίζεται ο αριθμός των αντικειμένων από τα οποία δεν έχουν εξαχθεί οι απαραίτητες πληροφορίες. Συγκεκριμένα, από τα συνολικά αποτελέσματα αφαιρείται ο startIndex, ο οποίος συγκρατεί τον αριθμό των δημοσιεύσεων που έχουν ανακτηθεί σε προηγούμενες αναζητήσεις, και το πλήθος των δημοσιεύσεων που αντλήθηκαν στην τρέχουσα επανάληψη. Σε περίπτωση που δεν υπάρχουν περαιτέρω αποτελέσματα, η αναζήτηση συνεχίζεται με το επόμενο πεδίο έως ότου διατρεχθεί ολόκληρη η λίστα με τα επιστημονικά πεδία. Η μέθοδος τερματίζει με την επιστροφή της λίστας των DOIs των δημοσιεύσεων που ανακτήθηκαν από το πρόγραμμα.

3.3.5 Άντληση metadata – Στάδιο 2

Για την ανάκτηση metadata δημοσιεύσεων, profiles αρθρογράφων και οργανισμών αξιοποιήθηκε το Python πακέτο pybliometrics που αναπτύχθηκε από τους Rose & Kitchin (2019) [39]. Το συγκεκριμένο πακέτο παρέχει μεταξύ άλλων API wrappers, τα οποία ενσωματώνονται στο API της Scopus για την απλή και εύκολη ανάκτηση δεδομένων οντοτήτων που σχετίζονται με την επιστημονική συγγραφή. Τα API wrappers που χρησιμοποιήθηκαν στο πλαίσιο της διπλωματικής είναι:

- AbstractRetrieval, για την άντληση metadata δημοσιεύσεων.
- AuthorRetrieval, για την άντληση metadata συγγραφέων.
- AbstractRetrieval, για την άντληση metadata οργανισμών.
- PlumXMetrics, για την άντληση αριθμού παραπομπών από ευρέως διαδεδομένες και αναγνωρισμένες βάσεις δεδομένων επιστημονικής βιβλιογραφίας όπως η CrossRef, η PubMed και η Scopus.

Οι παραπάνω κλάσεις δέχονται σαν παράμετρο εισόδου το αναγνωριστικό ID της Scopus (ή το DOI στην περίπτωση της AbstractRetrieval) προκειμένου να επιστρέψουν ένα αντικείμενο το οποίο θα περιέχει την βιβλιομετρική πληροφορία ως χαρακτηριστικά της αντίστοιχης οντότητας.

Κατά αντιστοιχία, όπως παρουσιάζεται στις Εικόνες 3.5, 3.6 και 3.7, αναπτύχθηκαν κλάσεις για δημοσιεύσεις, συγγραφείς και οργανισμούς με σκοπό την προ-επεξεργασία και την αποτελεσματικότερη διαχείριση της αντλούμενης βιβλιομετρικής πληροφορίας.

```
class Publication:
    def __init__(self, publicationInfo, year, doi):
        self.id = str(uuid.uuid4())
        self.doi = doi
        self.year = year
        self.title = applySqlSyntax(getSafeAttribute(publicationInfo, 'title', 'string'))
        self.journal = applySqlSyntax(getSafeAttribute(publicationInfo, 'publicationName', 'string'))
        self.abstract = applySqlSyntax(
            Publication.getAbstract(
                getSafeAttribute(publicationInfo, 'abstract', 'string'),
                getSafeAttribute(publicationInfo, 'description', 'string')
            )
        )
        self.keywords = applySqlSyntax(Publication.getKeywords(getSafeAttribute(publicationInfo, 'authkeywords', 'string')))
        self.fields = applySqlSyntax(Publication.getFields(getSafeAttribute(publicationInfo, 'subject_areas', 'string')))
        self.citationsCount = Publication.getMaximumCitationsCount(
            getSafeAttribute(publicationInfo, 'citedby_count', 'number'),
            doi
        )
        self.authorsNumber = Publication.getAuthorsNumber(getSafeAttribute(publicationInfo, 'authors', 'list'))
        self.affiliationsNumber = Publication.getAffiliationsNumber(getSafeAttribute(publicationInfo, 'affiliation', 'list'))
```

Εικόνα 3.5: Python κλάση για την διαχείριση metadata δημοσιεύσεων (πηγή: ιδία χρήση).

```
class Author:
    def __init__(self, authorInfo):
        self.id = str(uuid.uuid4())
        self.scopusId = str(getSafeAttribute(authorInfo, 'identifier', 'string'))
        self.orcidId = getSafeAttribute(authorInfo, 'orcid', 'string')
        self.firstName = applySqlSyntax(getSafeAttribute(authorInfo, 'given_name', 'string'))
        self.lastName = applySqlSyntax(getSafeAttribute(authorInfo, 'surname', 'string'))
        self.hIndex = getSafeAttribute(authorInfo, 'h_index', 'number')
        self.fieldsOfStudy = applySqlSyntax(Author.getFields(getSafeAttribute(authorInfo, 'subject_areas', 'string')))
        self.citationsCount = getSafeAttribute(authorInfo, 'cited_by_count', 'number')
        self.affiliations = applySqlSyntax(Author.getAffiliations(getSafeAttribute(authorInfo, 'affiliation_history', 'string')))
```

Εικόνα 3.6: Python κλάση για την διαχείριση metadata συγγραφέων (πηγή: ιδία χρήση).

```
class Organization:
    def __init__(self, organizationInfo):
        self.id = str(uuid.uuid4())
        self.scopusId = str(getSafeAttribute(organizationInfo, 'identifier', 'string'))
        self.name = applySqlSyntax(getSafeAttribute(organizationInfo, 'affiliation_name', 'string'))
        self.type1, self.type2 = Organization.getAffiliationTypes(organizationInfo)
        self.address = applySqlSyntax(getSafeAttribute(organizationInfo, 'address', 'string'))
        self.city = applySqlSyntax(getSafeAttribute(organizationInfo, 'city', 'string'))
        self.country = applySqlSyntax(getSafeAttribute(organizationInfo, 'country', 'string'))
```

Εικόνα 3.7: Python κλάση για την διαχείριση metadata οργανισμών (πηγή: ιδία χρήση).

Όλες οι παραπάνω κλάσεις δέχονται ως παράμετρο εισόδου τα αντικείμενα που παράγονται από τις κλάσεις AbstractRetrieval, AuthorRetrieval και AffiliationRetrieval (publicationInfo, authorInfo και organizationInfo αντίστοιχα). Στην περίπτωση της κλάσης Publication εισάγεται και το έτος δημοσίευσης που δόθηκε από τον χρήστη και το αναγνωριστικό DOI που αντλήθηκε με την ολοκλήρωση του Σταδίου 1 της διαδικασίας της εξαγωγής. Τα αναγνωριστικά Scopus ID που είναι απαραίτητα για τη χρήση των AuthorRetrieval και AffiliationRetrieval, είναι δυνατό να αποκτηθούν από το αντικείμενο εξόδου της AbstractRetrieval, καθώς περιέχει τα αναγνωριστικά των συγγραφέων και των οργανισμών που διασυνδέονται με τη συγγραφή μιας συγκεκριμένης δημοσίευσης.

Για την υλοποίηση των κλάσεων που αναφέρθηκαν παραπάνω, αναπτύχθηκαν μέθοδοι προ-επεξεργασίας των πρωτογενών δεδομένων, έτσι ώστε να αποφευχθούν σφάλματα τόσο στην διαχείριση τους όσο και στην μετέπειτα εισαγωγή τους στην βάση δεδομένων. Συγκεκριμένα, οι κοινές για όλες τις κλάσεις μέθοδοι που αναπτύχθηκαν και των οποίων έγινε χρήση είναι οι:

- `getSafeAttribute`, που πραγματοποιεί έλεγχο αν το εκάστοτε χαρακτηριστικό που επιχειρείται να αναγνωσθεί, υπάρχει όντως στο συγκεκριμένο αντικείμενο.
- `applySqlSyntax`, που είναι υπεύθυνη για την εξασφάλιση format, σε περιπτώσεις που το χαρακτηριστικό είναι string, το οποίο να είναι αποδεκτό από τους κανόνες σύνταξης του RDMBS που χρησιμοποιήθηκε.
- `getAbstract`, η οποία ελέγχει ποιο από τα χαρακτηριστικά Abstract η Description περιέχεται σε κάθε αντικείμενο δημοσίευσης και το αναθέτει σαν περίληψη του αντικειμένου.
- `getKeywords` και `getFields`, για την άντληση λέξεων – κλειδιών και επιστημονικών πεδίων αντίστοιχα. Τα συγκεκριμένα χαρακτηριστικά αντλούνται υπό την μορφή λιστών από την Scopus, οπότε σκοπός αυτών των μεθόδων είναι η μετατροπή τους σε strings.
- `getMaximumCitationsCount`, η οποία αξιοποιεί την κλάση PlumXMetrics του pybliometrics για τον μέγιστο αριθμό αναφορών της κάθε δημοσίευσης που παρέχεται από τις διαφορετικές βάσεις δεδομένων.
- `getAuthorsNumber` και `getAffiliationsNumber`, οι οποίες στοχεύουν στην εξαγωγή του αριθμού συγγραφέων και οργανισμών από την επεξεργασία των πρωτογενών δεδομένων της κάθε δημοσίευσης.
- `getAffiliationType`, που είναι υπεύθυνη για την κατηγοριοποίηση των οργανισμών βάσει των Τύπων 1 και 2 που αναφέρθηκαν στην [Παράγραφο 3.2.2](#). Για τον σκοπό αυτό, αξιοποιείται η κατηγορία στην οποία εντάσσεται ο εκάστοτε οργανισμός στους καταλόγους της Scopus. Αν ωστόσο αυτή δεν παρέχεται από τα αντλούμενα δεδομένα, η μέθοδος διατρέχει το όνομα του κάθε οργανισμού για συγκεκριμένες λέξεις – κλειδιά βάσει των οποίων και γίνεται τελικά η ένταξη σε κάποιον από τους καθορισμένους Τύπους.

3.4 Backend ανάπτυξη και σχεδιασμός API

3.4.1 Εργαλεία που χρησιμοποιήθηκαν

Η backend ανάπτυξη στην παρούσα εργασία εστιάζει στην ανάπτυξη ενός API που λαμβάνει τα αιτήματα των χρηστών και βάσει αυτών επιστρέφει τα κατάλληλα metadata. Συγκεκριμένα, αναπτύχθηκε κώδικας Python για την υλοποίηση endpoint το οποίο εκτελεί πολλαπλές λειτουργίες, όπως η ανάκτηση και η προ-επεξεργασία των πρωτογενών δεδομένων, η κατάλληλη καταχώρηση τους στη βάση δεδομένων, ο εντοπισμός διπλοτύπων και τέλος η επιστροφή όλης της απαραίτητης βιβλιογραφικής πληροφορίας στον χρήστη.



Για την ανάπτυξη του API αλλά και για την υλοποίηση όλων των μεθόδων που χρησιμοποιεί κατά την λειτουργία, χρησιμοποιήθηκαν τα Python πακέτα:

- Flask, το οποίο περιλαμβάνει τις απαραίτητες μεθόδους για την ανάπτυξη και την λειτουργία online sever με τη δυνατότητα αποστολής, λήψης και επεξεργασίας HTTP αιτημάτων.
- mysql, που καθιστά δυνατή την αλληλεπίδραση του API με την MySQL βάση δεδομένων για την καταχώρηση και ανάκτηση δεδομένων.
- fuzzywuzzy, το οποίο διευκόλυνε την διαδικασία αναγνώρισης διπλοτύπων μέσω της παροχής αλγορίθμων σύγκρισης strings.

3.4.2 Καταχώρηση metadata στη βάση δεδομένων

Στην Εικόνα 3.8 παρουσιάζεται ο ορισμός του endpoint αναζήτησης, με επίθεμα /search και με την GET ως αποδεκτή HTTP μέθοδο. Πρώτη ενέργεια που πραγματοποιείται από το συγκεκριμένο endpoint είναι η ανάκτηση των κριτηρίων αναζήτησης (λέξεις – κλειδιά, επιστημονικά πεδία, χρονικό εύρος κτλ.) από τα HTTP queries που λαμβάνει. Στη συνέχεια εγκαθιδρύει επικοινωνία με τη βάση δεδομένων μέσω της μεθόδου connect, η οποία επιστρέφει τους όρους cursor και connection. Οι όροι αυτοί χρησιμοποιούνται για την εκτέλεση και επικύρωση αντίστοιχα SQL queries. Στη συνέχεια, οι παραπάνω μεταβλητές εισάγονται στην μέθοδο extractMetadata, η οποία ξεκινά και τη διαδικασία εξαγωγής δεδομένων από τους καταλόγους της Scopus για κάθε έτος ξεχωριστά.

```

app = Flask(__name__)
CORS(app)

@app.route('/search', methods=['GET'])
def search():
    try:
        keywords = request.args.getlist("keywords")[0].split(',')
        booleans = request.args.getlist("booleans")[0].split(',')
        fields = request.args.getlist("fields")[0].split(',')
        year1 = request.args.get("year1")
        year2 = request.args.get("year2")
        scopusApiKey = request.args.get("scopusApiKey")

        connection, cursor = connect()

        for year in range(int(year1), int(year2)+1):
            extractMetadata(keywords, year, fields, booleans, scopusApiKey, connection, cursor)

```

Εικόνα 3.8: Ανάκτηση κριτηρίων αναζήτησης και έναρξη διαδικασίας εξαγωγής metadata (πηγή: ίδια χρήση).

Με την εκτέλεση της μεθόδου `extractMetadata`, αρχικά πραγματοποιείται η ανάκτηση μέσω της μεθόδου `getColumnLength` των μεγεθών των στηλών της βάσης δεδομένων που περιέχουν strings, όπως παρατηρείται στην Εικόνα 3.9. Αυτό εξυπηρετεί, όπως θα δειχθεί παρακάτω, στη δυναμική αύξηση του μεγέθους της εκάστοτε στήλης σε περίπτωση που κάποιο νέο προς εισαγωγή αντικείμενο είναι μεγαλύτερο από το τρέχων μέγεθος, έτσι ώστε να αποφευχθεί η απόρριψη του.

```

# getting columns size
doiLength = getColumnLength('DOI', 'scopus_publications', cursor)
titleLabelLength = getColumnLength('Title', 'scopus_publications', cursor)
journalLength = getColumnLength('Title', 'scopus_publications', cursor)
abstractLength = getColumnLength('Abstract', 'scopus_publications', cursor)
keywordsLength = getColumnLength('Keywords', 'scopus_publications', cursor)
fieldsLength = getColumnLength('Fields', 'scopus_publications', cursor)
fieldsOfStudyLength = getColumnLength('Fields_Of_Study', 'scopus_authors', cursor)
affiliationsLength = getColumnLength('Affiliations', 'scopus_authors', cursor)
affilNameLength = getColumnLength('Name', 'scopus_organizations', cursor)
affilAddressLength = getColumnLength('Address', 'scopus_publications', cursor)
affilCityLength = getColumnLength('City', 'scopus_organizations', cursor)
affilCountryLength = getColumnLength('Country', 'scopus_organizations', cursor)

```

Εικόνα 3.9: Ανάκτηση μεγεθών στηλών που περιέχουν strings (πηγή: ίδια χρήση).

Η διαδικασία συνεχίζεται με την εκτέλεση του κώδικα στην Εικόνα 3.10, όπου πρωτίστως ανακτώνται τα αναγνωριστικά DOIs μέσω της μεθόδου `getDois` που περιεγράφηκε στην Παράγραφο 3.3.4. Έπειτα, τα DOIs αξιοποιούνται για την εξαγωγή του αντικειμένου `publicationInfo` από την κλάση `AbstractRetrieval`, το οποίο ακολούθως εισάγεται στην κλάση `Publication`. Εκεί πραγματοποιείται η αρχική επεξεργασία των

δεδομένων, η οποία εξασφαλίζει την απαλοιφή null τιμών και την εφαρμογή αποδεκτής SQL σύνταξης στα αντίστοιχα strings κατά τον τρόπο που περιεγράφηκε στην Παράγραφο 3.3.5. Τελικά παράγεται το αντικείμενο publicationObj από την κλάση Publication, το οποίο αξιοποιείται στη συνέχεια για την εισαγωγή των metadata δημοσιεύσεων, που αντιστοιχούν στα χαρακτηριστικά του, στην βάση δεδομένων. Αν η εισαγωγή των χαρακτηριστικών του αντικειμένου διεξαχθεί με επιτυχία, η ένδειξη errorCodePub αποκτά τον κωδικό σφάλματος 0. Σε περίπτωση όμως που προκύψει κάποιο σφάλμα είτε κατά τη διαδικασία της άντλησης είτε κατά της διαδικασίας της καταχώρησης, η εκτέλεση του κώδικα περνά στον διαχειριστή σφαλμάτων που προβάλλεται στην Εικόνα 3.11.

```
# retrieving DOIs
dois = getDois(keywords, yearPublished, scopusFields, booleans, apiKey)

# getting publication metadata
print("Getting publications metadata ...")
for doi in tqdm(dois):
    try:
        publicationInfo = AbstractRetrieval(doi, view="FULL")
        publicationObj = Publication(publicationInfo, yearPublished, doi)

        while True:
            try:
                query = f"INSERT INTO scopus_publications VALUES('{publicationObj.id}', \
                    '{publicationObj.doi}', '{publicationObj.year}', '{publicationObj.title}', \
                    '{publicationObj.journal}', '{publicationObj.abstract}', '{publicationObj.keywords}', \
                    '{publicationObj.fields}', '{publicationObj.citationsCount}', \
                    '{publicationObj.authorsNumber}', '{publicationObj.affiliationsNumber}');"
                cursor.execute(query)
                connection.commit()
                errorCodePub = 0
                break
            except:
```

Εικόνα 3.10: Άντληση metadata δημοσιεύσεων και εισαγωγή τους στη βάση δεδομένων (πηγή: ίδια χρήση).

```
except Exception as err:
    if "Duplicate entry" not in str(err):
        errorCodePub = 2

    if "Data too long" in str(err):
        if "DOI" in str(err):
            try:
                query = f"ALTER TABLE scopus_publications MODIFY COLUMN DOI VARCHAR({doiLength});"
                cursor.execute(query)
                cursor.commit()
            except:
                pass
```

Εικόνα 3.11: Διαχείριση σφαλμάτων κατά την καταχώρηση των metadata στο RDBMS (πηγή: ίδια χρήση).

Στον διαχειριστή σφαλμάτων γίνεται αρχικά μια βασική διάκριση των πιθανών σφαλμάτων και αυτή είναι αν το σφάλμα προέκυψε λόγω της απόρριψης διπλοεγγραφής από

το RDBMS (ενέργεια που είναι επιθυμητή) ή λόγω κάποιας άλλης αιτίας. Στην πρώτη περίπτωση η τιμή της ένδειξης `errorCodePub` τίθεται ίση με 1 ενώ στη δεύτερη ίση με 2. Στη δεύτερη περίπτωση, πραγματοποιείται ένας επιπλέον έλεγχος με στόχο να εξακριβωθεί αν η διαδικασία εισαγωγής απέτυχε λόγω του μεγάλου μεγέθους κάποιων εκ των χαρακτηριστικών σε σχέση με το μέγεθος της αντίστοιχης στήλης στη βάση δεδομένων. Αν αυτό επαληθευτεί, το μέγεθος της επικείμενης στήλης αρχίζει να αυξάνεται μέσω `while loops`, έως ότου το αντικείμενο εισαχθεί επιτυχώς στην βάση. Στην εικόνα παραπάνω προβάλλεται ο κώδικας αύξησης μεγέθους της στήλης που περιέχει τα DOIs των papers, αλλά αντίστοιχα τμήματα κώδικα υλοποιούν την αυτόματη αναπροσαρμογή και των υπόλοιπων στηλών που περιέχουν strings.

Όπως φαίνεται στην Εικόνα 3.12, στην συνέχεια ελέγχεται η ένδειξη σφάλματος `errorCodePub` και αν αυτή υποδεικνύει πως δεν προέκυψε κάποιο σφάλμα κατά την διαχείριση των δεδομένων της τελευταίας δημοσίευσης (τιμή 0), η διαδικασία συνεχίζει με την άντληση και την καταχώρηση metadata συγγραφέων αυτής της δημοσίευσης, κατά παρόμοιο τρόπο με αυτόν που περιεγράφηκε για τα metadata δημοσιεύσεων. Ως επιπλέον βήμα παρεμβάλλεται η εξαγωγή της λίστας των συγγραφέων κάθε paper μέσω της κλάσης `AbstractRetrieval`, από την οποία απομονώνεται το Scopus ID (το πρώτο στοιχείο της λίστας `author` που αντιστοιχεί στο κάθε profile) για να εισαχθεί στην κλάση `AuthorRetrieval`.

```
if (errorCodePub == 0):
    try:
        authors = AbstractRetrieval(doi).authors

        for author in authors:
            authorInfo = AuthorRetrieval(author[0])
            authorObj = Author(authorInfo)

            while True:
                try:
                    query = f"INSERT INTO scopus_authors VALUES('{authorObj.id}',\
                        '{authorObj.scopusId}', '{authorObj.orcidId}', '{authorObj.firstName}',\
                        '{authorObj.lastName}', '{authorObj.fieldsOfStudy}', '{authorObj.affiliations}',\
                        '{authorObj.hIndex}', '{authorObj.citationsCount}');"
                    cursor.execute(query)
                    connection.commit()

                    errorCodeAut = 0
                    break
```

Εικόνα 3.12: Άντληση metadata συγγραφέων και εισαγωγή τους στη βάση δεδομένων (πηγή: ίδια χρήση).

Με το πέρας της διαδικασίας, ελέγχεται και η αντίστοιχη ένδειξη σφάλματος όπως παρατηρείται στην Εικόνα 3.13 προκειμένου να αρχίσει η εισαγωγή των IDs δημοσιεύσεων και συγγραφέων στον αντίστοιχο πίνακα συσχέτισης. Για να πραγματοποιηθεί αυτή η

ενέργεια, θα πρέπει είτε τα δεδομένα του τρέχοντος profile συγγραφέα να έχουν καταχωρηθεί με επιτυχία (κωδικός σφάλματος 0) είτε να έχουν απορριφθεί ως set δεδομένων που να έχει ήδη καταχωρηθεί (κωδικός σφάλματος 1). Δηλαδή, η βιβλιογραφική πληροφορία που είναι διασυνδεδεμένη με ένα συγκεκριμένο profile καταχωρείται μόνο μία φορά, ενώ το profile είναι πιθανό να σχετίζεται με την δημοσίευση πολλών και διαφορετικών άρθρων. Εφόσον λοιπόν, όλες οι συσχετίσεις μεταξύ δημοσιεύσεων και συγγραφέων είναι απαραίτητες για την πληρότητα των εξαγόμενων δεδομένων, στους αποδεκτούς κωδικού σφάλματος στην συγκεκριμένη περίπτωση εντάσσεται και ο κωδικός 1.

```
if (errorCodeAut in [0, 1]):
    query = f"INSERT INTO scopus_publications_authors VALUES('{publicationsScopusIds[doi]}',\
    '{scopusAuthors[authorObj.scopusId]}');"
    cursor.execute(query)
    connection.commit()

if (errorCodeAut == 0):
    try:
        authors = AbstractRetrieval(doi).authors
        for author in authors:
            affiliations = getAffiliationsIds(author[4])

            if (affiliations != "-"):
                authorId = str(AuthorRetrieval(author[0]).identifier)

                for affil in affiliations:
                    affiliationInfo = AffiliationRetrieval(int(affil), view="STANDARD")
                    organizationObj = Organization(affiliationInfo)

                    while True:
                        try:
                            query = f"INSERT INTO scopus_organizations VALUES('{organizationObj.id}','\
                            '{organizationObj.scopusId}','{organizationObj.name}','{organizationObj.type1}','\
                            '{organizationObj.type2}','{organizationObj.address}','{organizationObj.city}','\
                            '{organizationObj.country}');"
                            cursor.execute(query)
                            connection.commit()

                            errorCodeOrg = 0
                            break
```

Εικόνα 3.13: Άντληση metadata οργανισμών και εισαγωγή τους στη βάση δεδομένων (πηγή: ιδία χρήση).

Στην περίπτωση που ο κωδικός σφάλματος είναι 0, ξεκινάει και η διαδικασία άντλησης/καταχώρησης metadata και για τους οργανισμούς που σχετίζονται με το κάθε profile συγγραφέα. Η ακολουθία διεργασιών είναι παρόμοια με τις αντίστοιχες που αναφέρθηκαν προηγουμένως, με την διαφορά ότι παρεμβάλλεται και η εξαγωγή της λίστας των οργανισμών που περιέχεται στα metadata κάθε συγγραφέα. Όλοι οι όροι της λίστας διατρέχονται, ώστε να εξαχθεί για κάθε έναν από αυτούς το Scopus ID και στη συνέχεια να εισαχθεί στην κλάση AffiliationRetrieval για την ανάκτηση metadata του κάθε οργανισμού και την αποθήκευσή τους στη βάση δεδομένων. Κατά ανάλογο τρόπο με την καταχώρηση δεδομένων στον πίνακα συσχέτισης δημοσιεύσεων - συγγραφέων, πραγματοποιείται και η

εισαγωγή αναγνωριστικών στους πίνακες συσχέτισης δημοσιεύσεων – οργανισμών και συγγραφέων - οργανισμών, όπως παρουσιάζεται στην Εικόνα 3.14. Και σε αυτήν την περίπτωση, οι κωδικοί σφάλματος που επιτρέπουν την συμπλήρωση των εν λόγω πινάκων με δεδομένα είναι 0 και 1.

```
if (errorCodeOrg in [0, 1]):
    try:
        query = f"INSERT INTO scopus_publications_organizations VALUES('{publicationsScopusIds[doi]}', \
            '{scopusAffiliationsIds[organizationObj.scopusId]}');"
        cursor.execute(query)
        connection.commit()

    except Exception as err:
        if "Duplicate entry" not in str(err):
            print(f"{BLUE}Affiliation Metadata Inserting Error Info:{RESET}\n"
                f"DOI: {doi}\n"
                f"Affiliation Scopus ID: {organizationObj.scopusId}\n"
                f"Error: {str(err)}")

    try:
        query = f"INSERT INTO scopus_authors_organizations VALUES('{scopusAuthorsIds[authorId]}', \
            '{scopusAffiliationsIds[organizationObj.scopusId]}', {yearPublished});"
        cursor.execute(query)
        connection.commit()

    except Exception as err:
        if "Duplicate entry" not in str(err):
            print(f"{BLUE}Affiliation Metadata Inserting Error Info:{RESET}\n"
                f"DOI: {doi}\n"
                f"Affiliation Scopus ID: {organizationObj.scopusId}\n"
                f"Error: {str(err)}")
```

Εικόνα 3.14: Εισαγωγή IDs δημοσιεύσεων, συγγραφέων και οργανισμών στους πίνακες συσχετίσεων (πηγή: ιδία χρήση).

3.4.3 Αναγνώριση διπλοτύπων

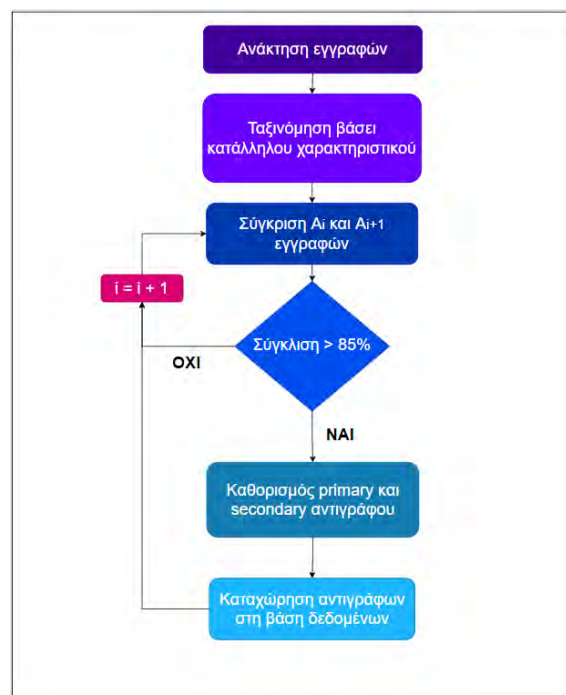
Την διαδικασία άντλησης και καταχώρησης των βιβλιομετρικών δεδομένων ακολουθεί η εφαρμογή μεθόδων εντοπισμού διπλοτύπων. Για κάθε οντότητα που σχετίζεται με την επιστημονική συγγραφή (δημοσιεύσεις, συγγραφείς και οργανισμοί) αναπτύχθηκε εξειδικευμένος αλγόριθμος αναγνώρισης πολλαπλών set metadata που αντιστοιχούν στην ίδια οντότητα. Πριν την ανάλυση των συγκεκριμένων αλγορίθμων, πρέπει να σημειωθεί πως αποφασίστηκε ο εντοπισμός διπλοτύπων να έπεται της καταχώρησης των δεδομένων στη SQL βάση για δύο κύριους λόγους.

Ο πρώτος εξ αυτών είναι η μείωση του όγκου δεδομένων που καλούνται να διαχειριστούν οι μέθοδοι αναγνώρισης. Θέτοντας περιορισμούς μοναδικού κλειδιού για αναγνωριστικά όπως DOIs και Scopus IDs, τα πολλαπλά αντίγραφα αντικειμένων απορρίπτονται αυτόματα από το RDBMS κάτι που ελαχιστοποιεί των όγκων δεδομένων που πρόκειται να επεξεργαστούν κατά την διαδικασία εντοπισμού διπλοτύπων. Σημειώνεται πως τα DOIs

επαρκώς αξιόπιστα αναγνωριστικά που χρησιμοποιούνται ευρέως για την αυθεντικοποίηση άρθρων και ερευνητών αντίστοιχα [40],[41],[42] ενώ τα Scopus IDs που αντιστοιχούν στην ίδια καταχώρηση συγχωνεύονται αυτόματα στο πιο επικαιροποιημένο Scopus ID μέσω των κλάσεων AbstractRetrieval, AuthorRetrieval και AffiliationRetrieval (Rose & Kitchin, 2019) [39].

Ο δεύτερος λόγος είναι η δραστική μείωση του αριθμού των απαραίτητων πράξεων σύγκρισης. Με την εξαγωγή των καταχωρήσεων από την βάση δεδομένων είναι δυνατό να εφαρμοστεί ταξινόμηση (sorting) βάσει ενός χαρακτηριστικού το οποίο παρουσιάζει σχετικά μικρές αλλαγές ανάμεσα στα πολλαπλά αντίγραφα. Μια τέτοιου είδους ταξινόμηση έχει σαν αποτέλεσμα τα διπλότυπα να συγκεντρώνονται σε groups, σε συνεχόμενες δηλαδή θέσεις. Με αυτόν τον τρόπο ελέγχεται κάθε καταχώρηση με την επόμενη, όπου στην περίπτωση η καταχωρήσεων αυτό οδηγεί στην τέλεση $n-1$ πράξεων σύγκρισης. Αντίθετα, αν τα αντίγραφα βρίσκονται διασκορπισμένα στην λίστα που ανακτάται, κάθε εγγραφή πρέπει να συγκριθεί με όλες τις υπόλοιπες (η A_n θα ελεγχθεί $n-1$ φορές, η A_{n-1} θα ελεγχθεί $n-2$ φορές κ.ο.κ) κάτι που οδηγεί σε $n^2/2$ πράξεις σύγκρισης, που είναι τάξης μεγέθους μεγαλύτερος αριθμός πράξεων από τον αντίστοιχο αριθμό στην περίπτωση εφαρμογής sorting.

Στην Εικόνα 3.15 παρουσιάζεται ο γενικός αλγόριθμος που εφαρμόζουν οι μέθοδοι που αναπτύχθηκαν για την ταυτοποίηση πολλαπλών αντιγράφων που αντιστοιχούν σε μια ενιαία οντότητα.



Εικόνα 3.15: Γενικός αλγόριθμος εντοπισμού διπλοτύπων (πηγή: ίδια χρήση).

Όπως προ – αναφέρθηκε, αρχικά ανακτώνται τα δεδομένα που καταχωρήθηκαν στη βάση δεδομένων, έπειτα από την άντληση τους από την Scopus. Κατά την ανάκτηση τους ταξινομούνται βάσει ενός κατάλληλου χαρακτηριστικού. Για τις δημοσιεύσεις κρίθηκε πως καταλληλότερο χαρακτηριστικό ταξινόμησης είναι ο τίτλος της κάθε δημοσίευσης, το ονοματεπώνυμο κάθε ερευνητή και το όνομα κάθε οργανισμού. Στη συνέχεια, πραγματοποιείται η σύγκριση μεταξύ διαδοχικών εγγραφών. Η διαδικασία της σύγκρισης πραγματοποιήθηκε μέσω της συνάρτησης ratio που παρέχεται από το Python πακέτο fuzzywuzzy. Η συγκεκριμένη συνάρτηση δέχεται ως παράμετρο εισόδου δύο strings τα οποία αντιστοιχούν σε συγκεκριμένα χαρακτηριστικά μεταξύ δύο καταχωρήσεων:

- Για τις δημοσιεύσεις, τέθηκαν ως κριτήρια σύγκρισης ο τίτλος, η περίληψη, οι λέξεις - κλειδιά και τα επιστημονικά πεδία της κάθε δημοσίευσης.
- Για τους συγγραφείς αρχικά συγκρίνονται τα ονόματα, τα επώνυμα, τα επιστημονικά πεδία και το ιστορικό διασυνδέσεων του κάθε συγγραφέα με οργανισμούς.
- Για τους οργανισμούς αξιοποιήθηκε το όνομα και η διεύθυνση του κάθε οργανισμού ως κριτήριο αναγνώρισης διπλοτύπων.

Στην περίπτωση που η ratio επιστρέψει ποσοστό σύγκλισης όλων των παραπάνω groups χαρακτηριστικών μεγαλύτερο του 85%, καθορίζεται το πρωτεύον (primary) και δευτερεύον (secondary) αντίγραφο και στη συνέχεια τα δύο αντίγραφα αποθηκεύονται στον αντίστοιχο πίνακα διπλοτύπων (δημοσιεύσεων, συγγραφέων ή οργανισμών αντίστοιχα) που έχει αναπτυχθεί στη βάση δεδομένων. Οι όροι primary και secondary χρησιμοποιούνται για να υποδείξουν το αντίγραφο που έχει τη μεγαλύτερη πιθανότητα να είναι το πιο έγκυρο ή αυτό με τη μεγαλύτερη βιβλιομετρική συνεισφορά.

Τα κριτήρια για αυτή την διάκριση ποικίλουν ανάλογα την οντότητα που εξετάζεται σε κάθε περίπτωση:

- Στην περίπτωση δημοσιεύσεων λαμβάνεται υπόψιν ο αριθμός παραπομπών του κάθε αντιγράφου.
- Στην περίπτωση συγγραφέων εξετάζεται ο αριθμός αναφορών του κάθε συγγραφέα.
- Στην περίπτωση οργανισμών ως πρωτεύον αντίγραφο ορίζεται το αντίγραφο με την πιο πρόσφατη ημερομηνία δημιουργίας profile στη Scopus.

Αν το ποσοστό σύγκλισης είναι μικρότερο του 85% για οποιοδήποτε από τα χαρακτηριστικά του κάθε set, η διαδικασία συνεχίζει την ανίχνευση με τον έλεγχο των δύο επόμενων καταχωρήσεων, έως ότου όλες εξ αυτών έχουν ελεγχθεί και έχουν αναγνωριστεί είτε ως πρωτότυπα είτε ως αντίγραφα (primary ή secondary).

3.4.4 Επιστροφή αποτελεσμάτων αναζήτησης

Αφού ολοκληρωθεί η καταχώρηση στη βάση δεδομένων των metadata που αντλήθηκαν από την πλατφόρμα της Scopus αλλά και των ζευγών καταγραφών που ταυτοποιήθηκαν ως πιθανά αντίγραφα, η διαδικασία της αναζήτησης αποπερατώνεται με την ανάκτηση δεδομένων από την SQL βάση δεδομένων και την αποστολή τους στον χρήστη.

Η διαδικασία της ανάκτησης των επιθυμητών δεδομένων από την βάση πραγματοποιείται μέσω της σύνταξης και εκτέλεσης κατάλληλου SQL query. Στις Εικόνες 3.16 και 3.17 προβάλλονται τα δύο τμήματα του query που χρησιμοποιείται για αυτόν τον σκοπό. Στο πρώτο τμήμα καθορίζονται τα πεδία των καταχωρήσεων που πρόκειται να επιστραφούν στον χρήστη μέσω της SQL εντολής SELECT. Πέραν αυτού όμως, διαμορφώνεται δυναμικά και ο πίνακας από τον οποίο θα ανακτηθούν οι καταχωρήσεις. Αυτό επιτυγχάνεται με τη δημιουργία ενώσεων INNER JOIN μεταξύ των πινάκων που περιέχουν τα metadata δημοσιεύσεων, συγγραφέων και οργανισμών και των πινάκων συσχέτισης μεταξύ αυτών των οντοτήτων.

```
basicQuery = f'SELECT scopus_publications.ID, scopus_publications.DOI, scopus_publications.Title, scopus_publications.Year,\n \n scopus_publications.Citations_Count, scopus_publications.Keywords, scopus_publications.Fields, scopus_authors.ID, \n \n scopus_authors.First_Name, scopus_authors.Last_Name, scopus_authors.Fields_Of_Study, scopus_authors.Citations_Count, \n \n scopus_authors.hIndex, scopus_organizations.ID, scopus_organizations.Name, scopus_organizations.Type_1, \n \n scopus_organizations.Type_2, scopus_organizations.City, scopus_organizations.Country \n \n FROM (((scopus_publications_authors \n \n INNER JOIN scopus_publications ON scopus_publications_authors.Publication_ID = scopus_publications.ID) \n \n INNER JOIN scopus_authors ON scopus_publications_authors.Author_ID = scopus_authors.ID) \n \n INNER JOIN scopus_authors_organizations ON scopus_authors.ID = scopus_authors_organizations.Author_ID) \n \n INNER JOIN scopus_organizations ON scopus_authors_organizations.Organization_ID = scopus_organizations.ID)\n'
```

Εικόνα 16: Τμήμα του SQL query όπου καθορίζονται τα επιθυμητά metadata (στήλες) για την ανάκτηση τους από την βάση (πηγή: ίδια χρήση).

Στο δεύτερο τμήμα του SQL query σχηματίζονται οι απαραίτητες συνθήκες για την ανάκτηση των επιθυμητών καταχωρήσεων σε κάθε αναζήτηση. Συγκεκριμένα, για κάθε παράμετρο που έχει εισάγει ο χρήστης σαν κριτήριο αναζήτησης και έχει παραληφθεί με τη μορφή λίστας από το API, δομείται και το αντίστοιχο string μέσω for loops:

- Σχηματισμός string για τον έλεγχο αν κάθε μια από τις λέξεις – κλειδιά συμπεριλαμβάνεται στον τίτλο, στην περίληψη ή στις αντίστοιχες λέξεις – κλειδιά του κάθε άρθρου. Επιπρόσθετα, στην δομή του συγκεκριμένου string συνεισφέρουν και οι τελεστές boolean προκειμένου να καθοριστεί η συσχέτιση μεταξύ των λέξεων – κλειδιών.
- Σχηματισμός string για να εξακριβωθεί αν μια καταχώρηση δημοσίευσης εντάσσεται μέσα στο χρονικό εύρος που έχει καθοριστεί.

- Σχηματισμός string για να ελεγχθεί αν το κάθε πεδίο που έχει δοθεί σαν παράμετρος εισόδου, περιλαμβάνεται στα αντίστοιχα επιστημονικά πεδία μιας δημοσίευσης.

```

conditionQuery = 'WHERE\n (\n'
# keywords
for i in range(len(keywords)):
    if (i == 0):
        tempQuery = f'scopus_publications.Keywords LIKE \'{{keywords[i].lower()}}%\n \
                        OR scopus_publications.Title LIKE \'{{keywords[i].lower()}}%\n \
                        OR scopus_publications.Abstract LIKE \'{{keywords[i].lower()}}%\n'
        conditionQuery = conditionQuery + tempQuery
    else:
        tempQuery = f'{{booleans[i-1]}} \'Keywords\' LIKE \'{{keywords[i].lower()}}%\n \
                        OR scopus_publications.Title LIKE \'{{keywords[i].lower()}}%\n \
                        OR scopus_publications.Abstract LIKE \'{{keywords[i].lower()}}%\n'
        conditionQuery = conditionQuery + tempQuery

# years range
conditionQuery = conditionQuery + ') \n AND \n'
conditionQuery = conditionQuery + f'scopus_publications.Year >= {{year1}}\n'
conditionQuery = conditionQuery + f'AND scopus_publications.Year <= {{year2}}\n AND \n (\n'

# fields
for i in range(len(fullNameFields)):
    if (i == 0):
        tempQuery = f'scopus_publications.Fields LIKE \'{{fullNameFields[i]}}%\n'
        conditionQuery = conditionQuery + tempQuery
    else:
        tempQuery = f'OR scopus_publications.Fields LIKE \'{{fullNameFields[i]}}%\n'
        conditionQuery = conditionQuery + tempQuery

```

Εικόνα 3.17: Σχηματισμός SQL query για την εισαγωγή συνθηκών προκειμένου να ανακτηθούν οι κατάλληλες καταχωρήσεις (πηγή: ιδία χρήση).

Τα δύο strings που αναπτύχθηκαν ενώνονται σε ένα ενιαίο query το οποίο και τελικά εκτελείται για την συλλογή των metadata και την εισαγωγή τους στο αντικείμενο απόκρισης (response object) της Εικόνας 3.18. Τα πεδία successful και hasResult αποτελούν ενδείξεις για το κατά πόσο προέκυψαν σφάλματα στη διαδικασία αναζήτησης και αν υπάρχουν αποτελέσματα προς επιστροφή αντίστοιχα. Στο πεδίο data ενσωματώνονται τα metadata που τελικά ανακτώνται από το API, ενώ μέσω του πεδίου 'variants' επιστρέφονται όλα τα αντίγραφα των καταχωρήσεων που περιέχονται στο αντικείμενο απόκρισης.

```

result = {
    "successful": "true", # or false
    "hasResult": "true", # or false
    "data": data,
    "variants": variants
}

```

Εικόνα 17: Αντικείμενο απόκρισης από την διαδικασία αναζήτησης (πηγή: ιδία χρήση).

Στην Εικόνα 3.19 προβάλλεται το δείγμα κώδικα για τον εντοπισμό διπλοτύπων δημοσιεύσεων εντός των αποτελεσμάτων αναζήτησης. Αφού ανακτηθούν όλα τα IDs των πιθανών αντιγράφων δημοσιεύσεων κατά το στάδιο της άντλησης, ελέγχεται αν και τα δύο αναγνωριστικά του κάθε ζεύγους αντιγράφων συναντώνται εντός της λίστας αναγνωριστικών δημοσιεύσεων που επιστρέφεται ως αποτέλεσμα αναζήτησης. Σε περίπτωση που αυτό διαπιστωθεί, το πρωτεύον αντίγραφο προστίθεται στην λίστα 'originals' και το δευτερεύον στη λίστα 'duplicates'. Πέραν των αναγνωριστικών των αντιγράφων, επιστρέφονται και κάποια επιπλέον metadata, όπως ο αριθμός παραπομπών στην περίπτωση των δημοσιεύσεων. Αυτό αποσκοπεί στην παράθεση κριτηρίων στον χρήστη για το ποιο αντίγραφο να διατηρηθεί και ποιο να απορριφθεί. Με αντίστοιχο τρόπο, εντάσσονται και τα δεδομένα αντιγράφων ερευνητών και οργανισμών στην λίστα 'variants' του αντικείμενο απόκρισης.

```
# checking publications duplicates
publicationsVariants = {
    "originals": [],
    "duplicates": []
}
#fetching original and duplicated copies
query = "SELECT * FROM scopus_publications_variants;"
cursor.execute(query)
fetchData = cursor.fetchall()
for pubVar in fetchData:
    originalId = pubVar[0]
    duplicateId = pubVar[1]
    # checking if both of a pair's variants are in the returned data
    if ((originalId in publicationsIds) & (duplicateId in publicationsIds)):
        subQuery = f"SELECT scopus_publications.Id, scopus_publications.Title, \
            scopus_publications.Citations_Count FROM scopus_publications \
            WHERE ID = '{originalId}';"
        cursor.execute(subQuery)
        variantData = cursor.fetchall()[0]
        publicationsVariants["originals"].append({
            "id": variantData[0],
            "title": variantData[1],
            "citationsCount": variantData[2]
        })

        subQuery = f"SELECT scopus_publications.Id, scopus_publications.Title, \
            scopus_publications.Citations_Count FROM scopus_publications \
            WHERE ID = '{duplicateId}';"
        cursor.execute(subQuery)
        variantData = cursor.fetchall()[0]
        publicationsVariants["duplicates"].append({
            "id": variantData[0],
            "title": variantData[1],
            "citationsCount": variantData[2]
        })
```

Εικόνα 18: Έλεγχος για ύπαρξη πιθανών αντιγράφων δημοσιεύσεων εντός των δεδομένων που επιστρέφονται (πηγή: ιδία χρήση).

3.5 Frontend ανάπτυξη και σχεδιασμός UI

3.5.1 Εργαλεία που χρησιμοποιήθηκαν

Για την ανάπτυξη του UI της πλατφόρμας χρησιμοποιήθηκε το JavaScript framework AngularJS το οποίο παρέχει τη δυνατότητα:

- Δημιουργίας HTML templates για την απεικόνιση γραφιστικών στοιχείων, την εισαγωγή παραμέτρων και την προβολή δεδομένων που ανακτώνται από το API.
- Δημιουργίας αρχείων TypeScript για την πραγματοποίηση HTTP calls και την διαχείριση των δεδομένων που επιστρέφονται μέσω αυτών.
- Data binding τεχνικών για την διασύνδεση δεδομένων μεταξύ των TypeScript μεταβλητών και των αντίστοιχων HTML στοιχείων.



3.5.2 Εισαγωγή και αποστολή στοιχείων αναζήτησης

Για την είσοδο των παραμέτρων αναζήτησης εκ μέρος των χρηστών αναπτύχθηκε φόρμα σε HTML κώδικα, η οποία περιέχει στοιχεία input:

- Τύπου 'text' για την εισαγωγή του Scopus API Key που διαθέτει ο χρήστης.
- Τύπου 'text' για την εισαγωγή λέξεων - κλειδίων.
- Τύπου 'select' για την επιλογή ενός εκ των τριών boolean τελεστών (AND, OR, AND NOT).
- Τύπου 'text' για την εισαγωγή του χρονικού εύρους σε έτη.
- Τύπου 'checkbox' για την επιλογή επιστημονικών πεδίων.

Μέσω της λειτουργίας data binding τα στοιχεία εισόδου της φόρμας ενσωματώνονται στην TypeScript μεταβλητή searchForm. Με την ολοκλήρωση της συμπλήρωσης των πεδίων εισόδου, οι τιμές τους εξάγονται από την μεταβλητή searchForm και εισάγονται ως HTTP παράμετροι στη μεταβλητή parameters. Ακολουθεί η αποστολή των παραμέτρων στην διεύθυνση URL του endpoint /search που περιεγράφηκε στην προηγούμενη ενότητα. Μέσω της μεθόδου subscribe, παρακολουθείται το συγκεκριμένο endpoint έως ότου ληφθεί κάποια απόκριση. Όταν αυτή ληφθεί από τη μέθοδο και σε περίπτωση που τα πεδία successful και hasResult έχουν τιμές 'true', τα metadata που περιέχονται στο πεδίο data και τα πιθανά αντίγραφα (αν υπάρχουν) του πεδίου 'variants' αναθέτονται στις αντίστοιχες μεταβλητές. Επίσης πραγματοποιείται έλεγχος για την ύπαρξη διπλοτύπων με την μεταβλητή duplicatesFlag να παίρνει την ανάλογη τιμή true ή false. Αυτό αποσκοπεί στην δυναμική ενεργοποίηση της λειτουργίας διαχείρισης αντιγράφων, μόνο σε περίπτωση που υπάρχουν. Η παραπάνω μέθοδος παρουσιάζεται στην Εικόνα 3.20 παρακάτω.


```

let parameters = new HttpParams();
parameters = parameters
.set('keywords', this.keywords)
.set('booleans', this.booleans)
.set('year1', this.year1)
.set('year2', this.year2)
.set('fields', this.fields)
.set('scopusApiKey', this.scopusApiKey)

this.http
.get(this.apiUrl, { params: parameters })
.subscribe((response: any) => {
  if (response.successful === "true" && response.hasResult === "true") {
    this.data = response['data'];
    this.variants = response['variants'];

    if (this.variants.publicationsVariants.duplicates.length > 0 ||
        this.variants.authorsVariants.duplicates.length > 0 ||
        this.variants.organizationsVariants.duplicates.length > 0) {
      this.duplicatesFlag = true;
    }
    else {
      this.duplicatesFlag = false;
    }
  }
});

```

Εικόνα 3.20: Αποστολή των παραμέτρων εισόδου στο API μέσω HTTP GET μεθόδου (πηγή: ιδία χρήση).

3.5.3 Προβολή και διαχείριση αποτελεσμάτων αναζήτησης

Με την ανάκτηση των δεδομένων από την απόκριση του API, είναι δυνατή στη συνέχεια η προβολή τους μέσω των κατάλληλων δομών και στοιχείων. Συγκεκριμένα, τα metadata των οντοτήτων επιστημονικής συγγραφής αναπαρίστανται μέσω πίνακα που παρέχεται από το build-in πακέτο Material της AngularJS. Η συγκεκριμένη δομή πίνακα προσφέρει ένα σύνολο από χρήσιμες λειτουργίες ανάγνωσης και οργάνωσης δεδομένων, όπως η δυνατότητα ταξινόμησης των σειρών του πίνακα βάσει κάποιου πεδίου – στήλης, η δημιουργία πολλαπλών σελίδων για την ευανάγνωστη προβολή των εγγραφών και το φιλτράρισμα των καταχωρήσεων βάσει επιθυμητών κριτηρίων.

Επιπρόσθετα, αναπτύχθηκαν και κάποιες λειτουργίες μέσω των οποίων οι χρήστες έχουν τη δυνατότητα:

- Λήψης των αποτελεσμάτων σε αρχείο Excel.
- Λήψης των αποτελεσμάτων σε αρχείο CSV.
- Διαχείρισης των πιθανών αντιγράφων που εντοπίστηκαν στα αποτελέσματα αναζήτησης.

Σημειώνεται πως σε περίπτωση εφαρμογής κάποιου φίλτρου, η εξαγωγή σε αρχεία Excel/CSV αφορά στοχευμένα δεδομένα, που ικανοποιούν τις συνθήκες του εκάστοτε φίλτρου. Κάθε αυτόν τον τρόπο παρέχεται μεγαλύτερη ευελιξία στους χρήστες κατά τη διαδικασία λήψης αποτελεσμάτων.

Με την λειτουργία διαχείρισης πιθανών αντιγράφων, πραγματοποιείται η μετάβαση του χρήστη σε ένα νέο παράθυρο όπου μπορεί να προβεί στις παρακάτω ενέργειες:

- Προβολή του αριθμού διπλοτύπων για δημοσιεύσεις, profiles ερευνητών και καταχωρήσεις οργανισμών.
- Πρόσβαση στη λίστα αντιγράφων της κάθε οντότητας ξεχωριστά, ώστε να εξετάσει τις εγγραφές που προτείνονται ως διπλότυπα.

Σε περίπτωση που η ύπαρξη πολλαπλών παραλλαγών μιας ενιαίας καταχώρησης εξακριβωθεί και από τον χρήστη, μπορεί στη συνέχεια να επιλέξει την καταχώρηση από κάθε ζεύγος που επιθυμεί να συμπεριλαμβάνεται στην τελική εξαγωγή δεδομένων, βάσει των βιβλιομετρικών δεδομένων που παρατίθενται για κάθε αντίγραφο.

Στην περίπτωση των δημοσιεύσεων αξιοποιούνται οι μέθοδοι `updatePublicationsVariantsList` της Εικόνας 3.21. Μέσω της συγκεκριμένης μεθόδου, το αναγνωριστικό του επιλεγμένου αντίγραφου κάθε ζεύγους διπλοτύπων είτε προστίθεται στην λίστα `publicationsToBeRemoved` είτε αφαιρείται από αυτήν αν προηγουμένως έχει προστεθεί στην λίστα (σε περίπτωση που ο χρήστης επιθυμεί να αλλάξει αντίγραφο προς διαγραφή).

```
updatePublicationsVariantsList(publicationId:string, allowDuplicates:boolean = false) {  
  let index = this.publicationsToBeRemoved.indexOf(publicationId);  
  if (index >= 0 && !allowDuplicates) {  
    this.publicationsToBeRemoved.splice(index,1);  
  }  
  else {  
    this.publicationsToBeRemoved.push(publicationId);  
  }  
}
```

Εικόνα 3.21: Μέθοδος για την συμπλήρωση της λίστας αντιγράφων προς διαγραφή (πηγή: ίδια χρήση).

Η μέθοδος `clearPublicationsDuplicates` της Εικόνας 3.22, αξιοποιεί τα αναγνωριστικά που αποθηκεύτηκαν στην λίστα `publicationsToBeRemoved` για να φιλτράρει τα τελικά δεδομένα από τις καταχωρήσεις στις οποίες αντιστοιχούν σε αυτά τα αναγνωριστικά. Αυτό επιτυγχάνεται μέσω του πρώτου set for loops. Μέσω του δεύτερου set for loops, τόσο τα

αντίγραφα που παρέμειναν όσο και αυτά που απορρίφθηκαν, διαγράφονται από τις λίστες primary και secondary αντίστοιχα, εφόσον πλέον δεν τίθεται θέμα διπλοτύπων για το συγκεκριμένο ζεύγος. Τέλος, τα δεδομένα του πίνακα ανανεώνονται με τα φιλτραρισμένα πλέον δεδομένα, ενώ και η λίστα publicationsToBeRemoved εκκαθαρίζεται. Αντίστοιχες μέθοδοι αναπτύχθηκαν για αντίγραφα άρθρων και οργανισμών.

```
clearPublicationsDuplicates() {
  let filteredData = this.data[0];

  // filter data from selected duplicates
  for(let i = 0; i < this.publicationsToBeRemoved.length; i++) {
    filteredData = filteredData.filter((x:any) => x.publicationId !== this.publicationsToBeRemoved[i]);
  }

  // remove each pair's original and duplicate
  // from primary and secondary variants lists accordingly
  for (let i = 0; i < this.publicationsToBeRemoved.length; i++) {
    let index = this.primaryPublications.data.indexOf(
      this.primaryPublications.data
        .find((x:any) => x.id === this.publicationsToBeRemoved[i])
    );
    ? this.primaryPublications.data.indexOf(
      this.primaryPublications.data
        .find((x:any) => x.id === this.publicationsToBeRemoved[i])
    )
    : this.secondaryPublications.data.indexOf(
      this.secondaryPublications.data
        .find((x:any) => x.id === this.publicationsToBeRemoved[i]);
    );

    this.primaryPublications.data.splice(index,1);
    this.secondaryPublications.data.splice(index,1);
  }

  // update data after filtering
  this._searchService.setTableData(filteredData);
  this.publicationsVariantsNumber = this.data[1].publicationsVariants.duplicates.length;
  this.publicationsToBeRemoved = [];
  this.selectMode('reset');
}
```

Εικόνα 3.22: Μέθοδος για την εκκαθάριση δεδομένων από διπλότυπα (πηγή: ίδια χρήση).

3.6 Σύνοψη κεφαλαίου

Σε αυτό το κεφάλαιο περιεγράφηκε η μεθοδολογία που ακολουθήθηκε και οδήγησε στην υλοποίηση της αρχιτεκτονικής που παρουσιάζεται στην Εικόνα 3.23. Στο πρώτο βήμα αποφασίστηκε πως θα χρησιμοποιηθεί το community server σύστημα MySQL ως RDBMS και ορίστηκε το σχήμα βάσης που θα φιλοξενηθεί σε αυτό. Το σχήμα περιλαμβάνει τρεις βασικές οντότητες - πίνακες (Δημοσιεύσεις, Συγγραφείς και Οργανισμοί) βάσει των οποίων οργανώνεται και η βιβλιομετρική πληροφορία και τρεις πίνακες συσχέτισης για την διασύνδεση των οντοτήτων. Εκτός των χαρακτηριστικών που αντλούνται άμεσα από πηγές δεδομένων, εντάχθηκαν στο σχήμα μετρικά (τύποι οργανισμών, πλήθος συγγραφέων/οργανισμών ανά δημοσίευση) που προέκυψαν από την επεξεργασία των

αντλούμενων δεδομένων, με στόχο την εκτίμηση της αποτελεσματικότητας του αλγόριθμου εξαγωγής που αναπτύσσεται και την παράθεση επιπλέον βιβλιομετρικής πληροφορίας.

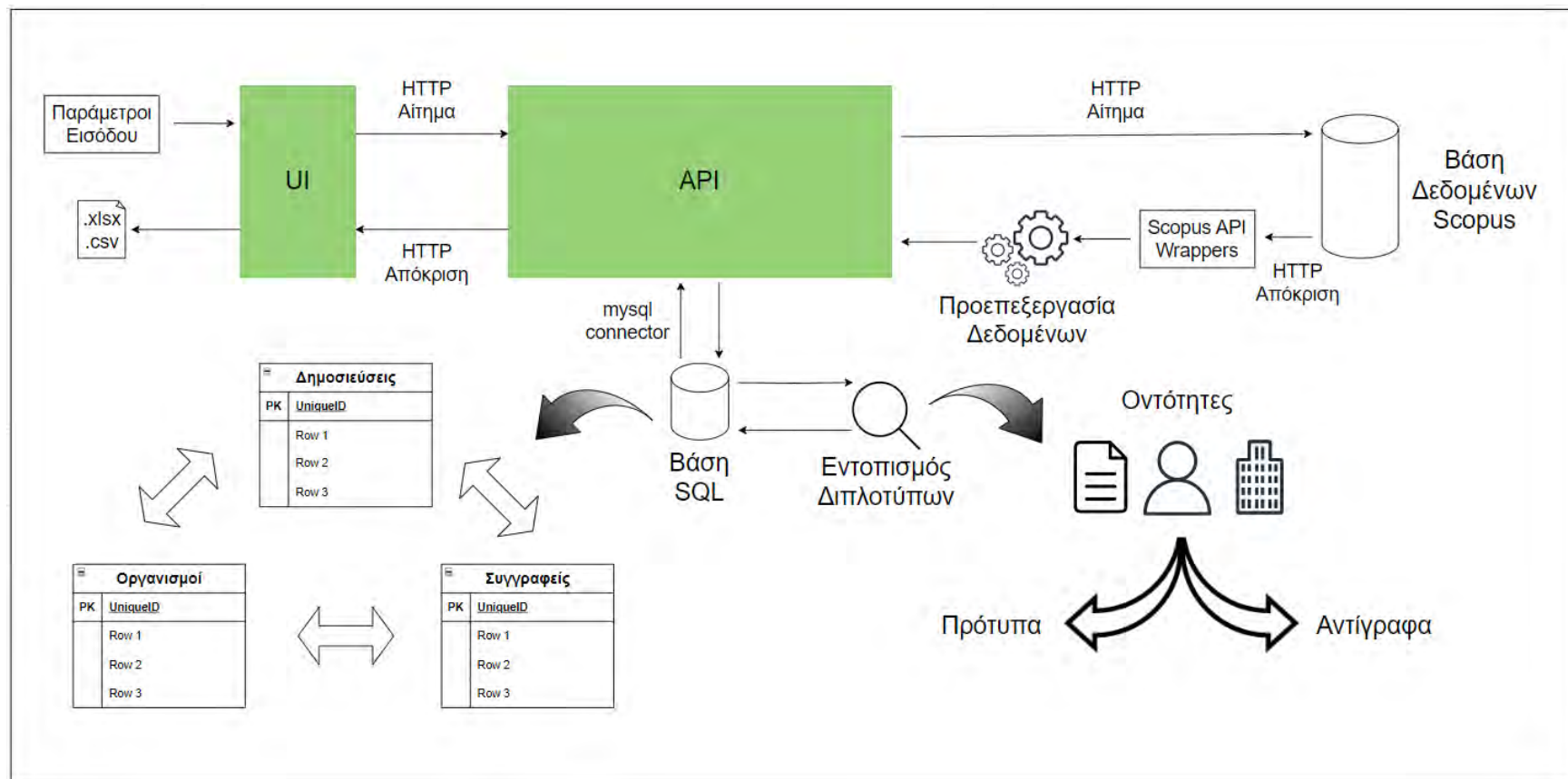
Έπειτα από διερεύνηση πολλαπλών πηγών δεδομένων, διαπιστώθηκε πως η πλατφόρμα της Scopus καλύπτει επαρκώς τις ανάγκες σε metadata της διπλωματικής εργασίας. Με την απόκτηση API Key για εξουσιοδοτημένη πρόσβαση στο περιεχόμενο της Scopus μέσω του API της, οριστικοποιήθηκε η χρήση της βάσης δεδομένων της Scopus ως πηγή metadata επιστημονικής βιβλιογραφίας.

Η διαδικασία της άντλησης δεδομένων από την Scopus διακριτοποιήθηκε σε δύο στάδια: στο πρώτο αναπτύσσεται μέθοδος σε κώδικα Python για την ανάκτηση DOIs δημοσιεύσεων και στο δεύτερο η αξιοποίηση αυτών των αναγνωριστικών για συλλογή metadata μέσω του Python πακέτου pybliometrics και των Scopus API wrappers που προσφέρει.

Οι μέθοδοι αυτές εκτελούνται από το API που αναπτύσσεται μέσω του Python πακέτου Flask. Την ανάκτηση των metadata ακολουθεί η οργάνωση και η προ – επεξεργασία τους (απαλοιφή null τιμών, εφαρμογή κατάλληλης SQL σύνταξης), με στόχο την ομαλή καταχώρηση των εξαγόμενων από την Scopus δεδομένων στην SQL βάση. Η αποθήκευση των δεδομένων βασίζεται στην εγκαθίδρυση HTTP επικοινωνίας μεταξύ του API και του MySQL server, την οποία προσφέρει το πακέτο mysql connector/

Η καταχώρηση των δεδομένων στη SQL βάση ενεργοποιεί τον μηχανισμό ανίχνευσης πιθανών καταγραφών που αντιστοιχούν στην ίδια οντότητα (άρθρο, ερευνητής ή οργανισμός). Ο μηχανισμός αυτός έχει υλοποιηθεί σε κώδικα Python και η λειτουργία του αφορά την σύγκριση καταχωρήσεων και την αποθήκευση τους στη SQL βάση, αν διαπιστωθεί πως υπάρχει σύγκλιση μεγαλύτερη του 85%. Το ποσοστό σύγκλισης βασίζεται στην σύγκριση συγκεκριμένων στοιχείων κειμένου μεταξύ πιθανών αντιγράφων, διαδικασία που επιτελείται μέσω κατάλληλων μεθόδων που παρέχει το Python πακέτο συναρτήσεων fuzzywuzzy. Με την ολοκλήρωση της διαδοχής των παραπάνω λειτουργιών, τα αποτελέσματα αναζήτησης (metadata και διπλότυπα) επιστρέφονται στους χρήστες.

Η αλληλοεπίδραση των χρηστών με την πλατφόρμα πραγματοποιείται μέσω του code – free UI που αναπτύσσεται, το οποίο δίνει τη δυνατότητα στον χρήστη να επιλέγει γρήγορα και εύκολα παραμέτρους αναζήτησης δημοσιεύσεων και να διαχειρίζεται ευέλικτα τα αποτελέσματα της κάθε αναζήτησης. Αυτό περιλαμβάνει την διαχείριση των προτεινόμενων από την πλατφόρμα διπλοτύπων, το φιλτράρισμα και την εξαγωγή σε αρχεία Excel και CSV των metadata που επεστράφησαν ως απόκριση από το API.



Εικόνα 3.23: Αρχιτεκτονική της πλατφόρμας (πηγή: ιδία χρήση).

Κεφάλαιο 4 ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

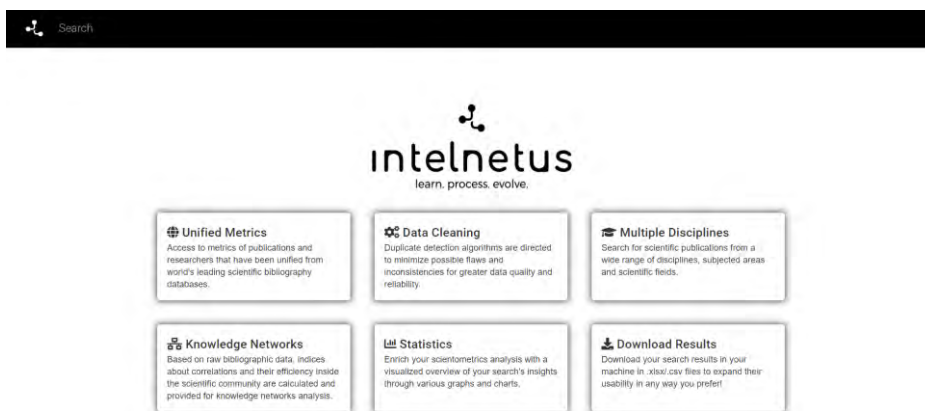
4.1 Εισαγωγή

Στο παρόν κεφάλαιο θα παρουσιαστούν τα αποτελέσματα της διπλωματικής εργασίας, στα οποία συμπεριλαμβάνονται το UI μέσω του οποίου αξιοποιούνται οι δυνατότητες της πλατφόρμας από τους χρήστες και τα metadata που επιστρέφονται έπειτα από την πραγματοποίηση διαφορετικών σεναρίων αναζήτησης. Επιπρόσθετα, υπολογίζονται και οι δείκτες αποτελεσματικότητας του αλγόριθμου εξαγωγής δεδομένων που αναφέρθηκαν στην Παράγραφο 3.2.2, προκειμένου να διαπιστωθεί αν υπήρξε διαφυγή δεδομένων κατά την άντληση τους από την πλατφόρμα της Scopus και σε τι ποσοστό. Με την παράθεση αλλά και την αξιολόγηση των παραπάνω αποτελεσμάτων, εξάγονται τελικά συμπεράσματα για την αποτελεσματικότητα και τους περιορισμούς της διαδικτυακής πλατφόρμας που αναπτύσσεται.

4.2 Παρουσίαση UI

4.2.1 Αρχική σελίδα

Στην Εικόνα 4.1 παρουσιάζεται η αρχική σελίδα της πλατφόρμας Intelnetus. Σε αυτήν περιλαμβάνεται το όνομα και το λογότυπο της πλατφόρμας, ενώ ένα σύνολο από διαδραστικές κάρτες χρησιμοποιείται για την προβολή των ήδη ενσωματωμένων λειτουργιών αλλά και των δυνατοτήτων που θα συμπεριληφθούν μελλοντικά στο λειτουργικό εύρος. Μέσω της μπάρας πλοήγησης στην κορυφή της σελίδας, είναι δυνατή η μετάβαση στην σελίδα αναζήτησης.



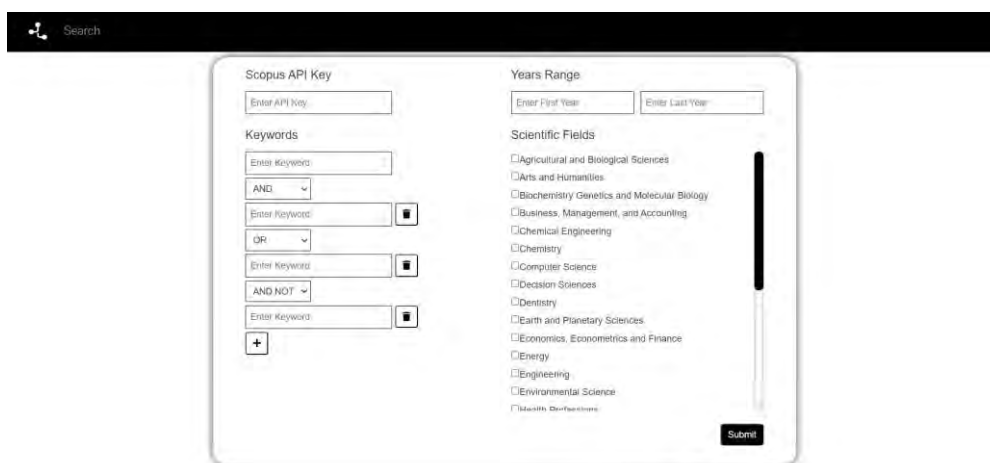
Εικόνα 4.1: Αρχική σελίδα της πλατφόρμας Intelnetus (πηγή: ιδία χρήση).

4.2.2 Σελίδα αναζήτησης

Στην Εικόνα 4.2 παρουσιάζεται η σελίδα αναζήτησης της πλατφόρμας, η οποία αποτελείται από μια δομή φόρμας. Στη συγκεκριμένη φόρμα εισάγονται όλες οι απαραίτητες παράμετροι μέσω των αντίστοιχων πεδίων, προκειμένου να ξεκινήσει η εκτέλεση της διαδικασίας της αναζήτησης:

- Το Scopus API Key που διαθέτει ο χρήστης.
- Λέξεις – κλειδιά και boolean τελεστές (AND, OR, AND NOT). Είναι εφικτή μάλιστα η πρόσθεση νέων πεδίων και τελεστών μέσω του αντίστοιχου κουμπιού. Ως προεπιλογή, οι λέξεις που θα εισαχθούν θα αναζητηθούν στους τίτλους, τις περιλήψεις και τις αντίστοιχες λέξεις κλειδιά των δημοσιεύσεων εντός της βάσης δεδομένων της Scopus.
- Χρονικό εύρος δημοσίευσης σε έτη.
- Επιστημονικά πεδία που συναντώνται στους καταλόγους δεδομένων της Scopus.

Με την συμπλήρωση όλων των παραπάνω πεδίων, ο χρήστης μπορεί να ξεκινήσει την διαδικασία αναζήτησης πατώντας το ενδεικτικό κουμπί ‘Submit’.

The image shows a web-based search interface for Scopus. At the top, there is a dark header with a magnifying glass icon and the word 'Search'. Below this, the main search area is divided into several sections. On the left, there is a 'Scopus API Key' section with a text input field. Below that is a 'Keywords' section with a text input field, a dropdown menu for boolean operators (currently showing 'AND'), and a list of input fields for keywords with a '+' button at the bottom. To the right of the keywords is a 'Years Range' section with two input fields for 'Enter First Year' and 'Enter Last Year'. Below the years range is a 'Scientific Fields' section with a list of checkboxes for various fields of study, including Agricultural and Biological Sciences, Arts and Humanities, Biochemistry Genetics and Molecular Biology, Business, Management, and Accounting, Chemical Engineering, Chemistry, Computer Science, Decision Sciences, Dentistry, Earth and Planetary Sciences, Economics, Econometrics and Finance, Energy, Engineering, Environmental Science, and Materials Science. A vertical scrollbar is visible on the right side of the scientific fields list. At the bottom right of the search area is a 'Submit' button.

Εικόνα 4.2: Σελίδα αναζήτησης της πλατφόρμας Intelnetus (πηγή: ιδία χρήση).

4.3 Σενάρια αναζήτησης

4.3.1 Σενάριο αναζήτησης 1 – Δοκιμαστική αναζήτηση

Το συγκεκριμένο σενάριο αναζήτησης αποσκοπεί στον έλεγχο της ομαλής λειτουργίας και της αποτελεσματικότητας των μεθόδων εξαγωγής και διαχείρισης metadata, εντός σύντομου σχετικά χρονικού διαστήματος για σκοπούς παρουσίασης των δυνατοτήτων της πλατφόρμας. Για τον σκοπό αυτό διερευνήθηκαν πολλαπλές περιπτώσεις παραμέτρων

εισόδου, ώστε να επιστραφεί ένας εύκολα διαχειρίσιμος αριθμός αποτελεσμάτων που να επιτρέπει την ολοκλήρωση της διαδικασίας σε χρονικό διάστημα μερικών λεπτών.

Τα κριτήρια αναζήτησης που τελικά επιλέχθηκαν για το συγκεκριμένο σενάριο αναζήτησης προβάλλονται στην Εικόνα 4.3 και είναι:

- ‘artificial intelligence’, ‘machine learning’, ‘deep learning’, ‘learning algorithm’ και ‘pattern recognition’ ως λέξεις – κλειδιά. Για την συσχέτιση των λέξεων – κλειδίων επιλέχθηκε ο τελεστής OR (αναζητούνται metadata δημοσιεύσεων που να περιέχουν τουλάχιστον μία από τις παραπάνω λέξεις στον τίτλο, την περίληψη ή τις λέξεις – κλειδιά τους).
- 2022 ως έτος δημοσίευσης (αρχικό και τελικό έτος ταυτίζονται).
- ‘social sciences’ ως επιστημονικό πεδίο.

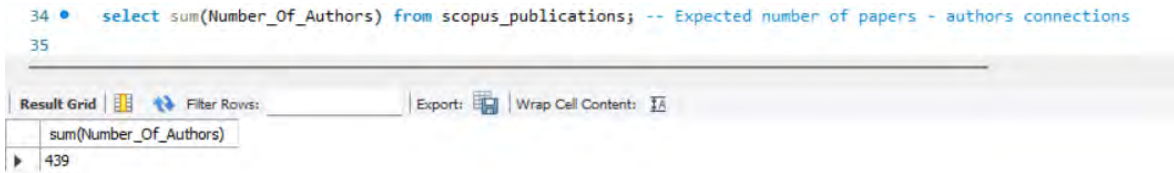
The image shows a web form for Scopus search. It includes a 'Scopus API Key' field with a placeholder 'Enter API Key'. A 'Years Range' section has two input boxes, both containing '2022'. The 'Keywords' section contains a list of search terms: 'artificial intelligence', 'machine learning', 'deep learning', 'learning algorithm', and 'pattern recognition', separated by 'OR' dropdowns. Each term has a trash icon to its right. A '+' button is at the bottom of the list. The 'Scientific Fields' section is a vertical list of checkboxes: Environmental Science, Health Professions, Immunology and Microbiology, Materials Science, Mathematics, Medicine, Multidisciplinary, Neuroscience, Nursing, Pharmacology, Toxicology, and Pharmaceutics, Physics and Astronomy, Psychology, Social Sciences (which is checked), and Veterinary. A vertical scrollbar is on the right of this list. A 'Submit' button is at the bottom right of the form.

Εικόνα 4.3: Παράμετροι εισόδου σεναρίου αναζήτησης 1 (πηγή: ιδία χρήση).

Τα παραπάνω κριτήρια αναζήτησης έχουν σαν αποτέλεσμα την ανάκτηση metadata 387 καταχωρήσεων. Ο αναμενόμενος αριθμός συνδέσεων μεταξύ άρθρων – συγγραφέων και άρθρων - οργανισμών υπολογίζεται μέσω του αθροίσματος των χαρακτηριστικών ‘Number Of Authors’ και ‘Number Of Affiliations’ αντίστοιχα για όλα τα papers που αντλήθηκαν. Αυτός υπολογίζεται μέσω των queries στις Εικόνες 4.4 και 4.5 στις 439 και 220 καταχωρήσεις αντίστοιχα. Ο πραγματικός αριθμός διασυνδέσεων άρθρων - συγγραφέων και άρθρων -οργανισμών που αποθηκεύτηκαν επιτυχώς στην SQL βάση υπολογίζεται μέσω των queries στις Εικόνες 4.6 και 4.7 στις 439 και 220 καταχωρήσεις επίσης. Συνεπώς, η διαφυγή

πληροφορίας διασύνδεσης για το δεδομένο set metadata δημοσιεύσεων που ανακτήθηκε είναι μηδαμινή, κάτι που αποτελεί θετική ένδειξη για την αποτελεσματικότητα του μηχανισμού εξαγωγής και διαχείρισης δεδομένων.

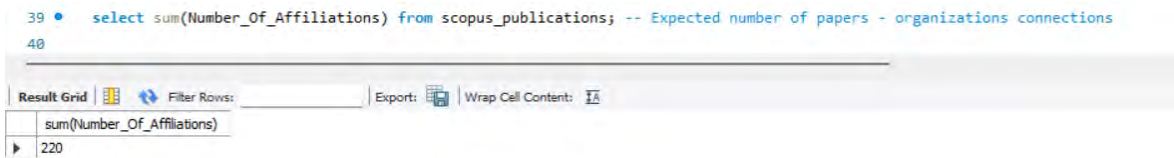
```
34 • select sum(Number_Of_Authors) from scopus_publications; -- Expected number of papers - authors connections
35
```



sum(Number_Of_Authors)
439

Εικόνα 4.4: Αναμενόμενος αριθμός διασυνδέσεων δημοσιεύσεων και συγγραφέων (πηγή: ίδια χρήση).

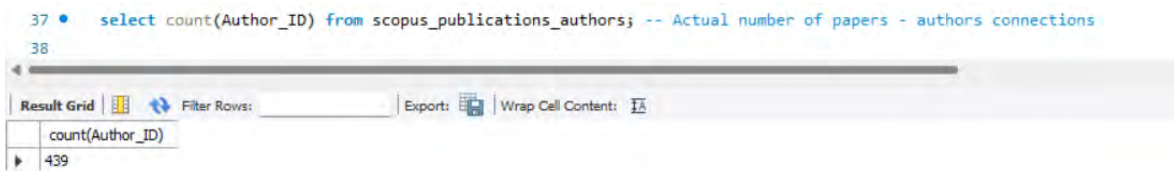
```
39 • select sum(Number_Of_Affiliations) from scopus_publications; -- Expected number of papers - organizations connections
40
```



sum(Number_Of_Affiliations)
220

Εικόνα 4.5: Αναμενόμενος αριθμός διασυνδέσεων δημοσιεύσεων και οργανισμών (πηγή: ίδια χρήση).

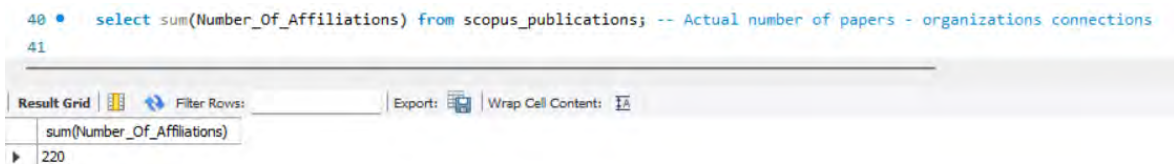
```
37 • select count(Author_ID) from scopus_publications_authors; -- Actual number of papers - authors connections
38
```



count(Author_ID)
439

Εικόνα 4.6: Πραγματικός αριθμός διασυνδέσεων δημοσιεύσεων και συγγραφέων (πηγή: ίδια χρήση).

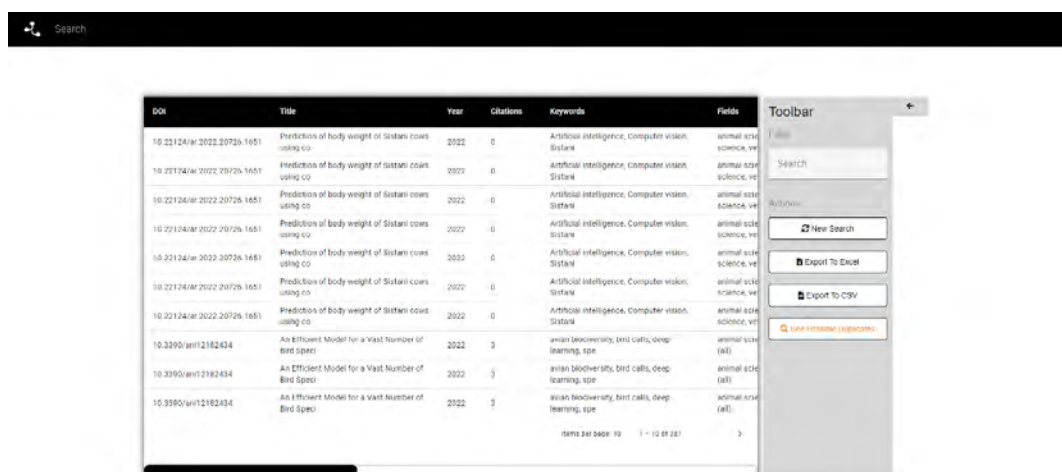
```
40 • select sum(Number_Of_Affiliations) from scopus_publications; -- Actual number of papers - organizations connections
41
```



sum(Number_Of_Affiliations)
220

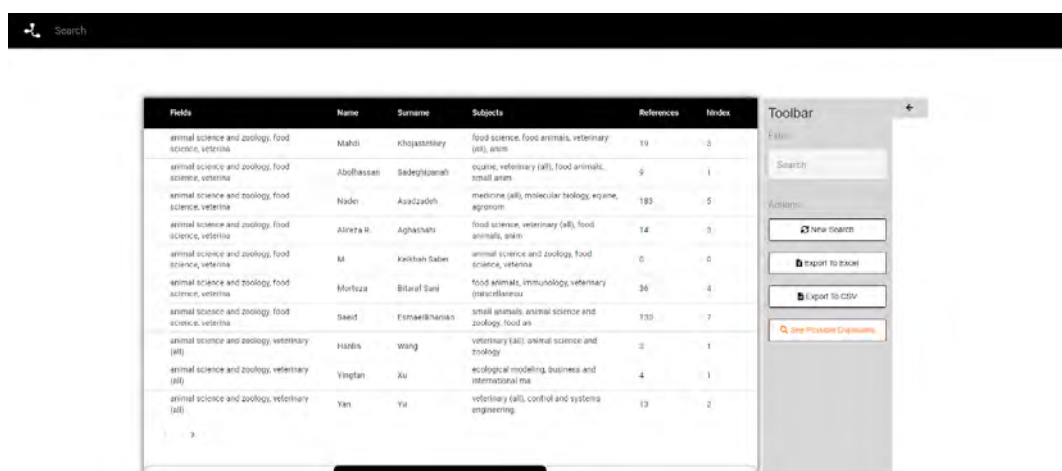
Εικόνα 4.7: Πραγματικός αριθμός διασυνδέσεων δημοσιεύσεων και οργανισμών (πηγή: ίδια χρήση).

Με την ολοκλήρωση της διαδικασίας αναζήτησης, τα metadata άρθρων, συγγραφέων και οργανισμών προβάλλονται μέσω του πίνακα συσχέτισης στις Εικόνες 4.8, 4.9 και 4.10. Η συσχέτιση αυτών των οντοτήτων επιτυγχάνεται με την συγκέντρωση σε μια ενιαία γραμμή των metadata του κάθε άρθρου, των profiles των ερευνητών – συγγραφέων, καθώς και των οργανισμών με τους οποίους διασυνδέονται οι ερευνητές κατά την συγγραφή του. Η καταχώρηση κάθε συγγραφέα επαναλαμβάνεται τόσες φορές όσες είναι και οι διασυνδέσεις του με οργανισμούς για το εκάστοτε άρθρο και αντίστοιχα, η καταχώρηση κάθε άρθρου επαναλαμβάνεται σε αριθμό ανάλογα του πλήθους των συγγραφέων του.



DOI	Title	Year	Citations	Keywords	Fields
10.22124/ir.2022.20726.1651	Prediction of body weight of Sistani cows using co	2022	0	Artificial intelligence, Computer vision, Sistani	animal science, veterinary
10.22124/ir.2022.20726.1651	Prediction of body weight of Sistani cows using co	2022	0	Artificial intelligence, Computer vision, Sistani	animal science, veterinary
10.22124/ir.2022.20726.1651	Prediction of body weight of Sistani cows using co	2022	0	Artificial intelligence, Computer vision, Sistani	animal science, veterinary
10.22124/ir.2022.20726.1651	Prediction of body weight of Sistani cows using co	2022	0	Artificial intelligence, Computer vision, Sistani	animal science, veterinary
10.22124/ir.2022.20726.1651	Prediction of body weight of Sistani cows using co	2022	0	Artificial intelligence, Computer vision, Sistani	animal science, veterinary
10.22124/ir.2022.20726.1651	Prediction of body weight of Sistani cows using co	2022	0	Artificial intelligence, Computer vision, Sistani	animal science, veterinary
10.22124/ir.2022.20726.1651	Prediction of body weight of Sistani cows using co	2022	0	Artificial intelligence, Computer vision, Sistani	animal science, veterinary
10.3390/ani12182434	An Efficient Model for a Vast Number of Bird Spec	2022	3	avian biodiversity, bird calls, deep learning, spe	animal science (all)
10.3390/ani12182434	An Efficient Model for a Vast Number of Bird Spec	2022	3	avian biodiversity, bird calls, deep learning, spe	animal science (all)
10.3390/ani12182434	An Efficient Model for a Vast Number of Bird Spec	2022	3	avian biodiversity, bird calls, deep learning, spe	animal science (all)

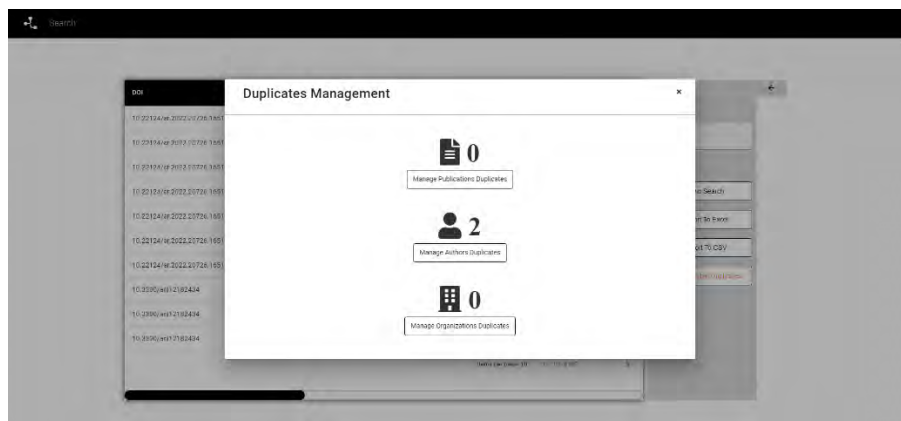
Εικόνα 4.8: Τμήμα πίνακα συσχέτισης με τα metadata δημοσιεύσεων (πηγή: ιδία χρήση).



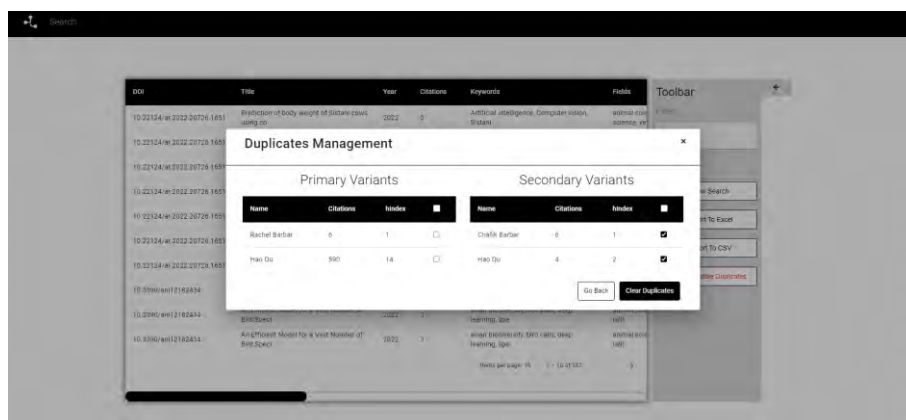
Fields	Name	Surname	Subjects	References	Index
animal science and zoology, food science, veterinary	Mahdi	Khojastehzadeh	food science, food animals, veterinary (all), zoon	19	0
animal science and zoology, food science, veterinary	Abdolhasan	Sadeghi-Javan	equine, veterinary (all), food animals, small anim	9	1
animal science and zoology, food science, veterinary	Nader	Azadzadeh	medicine (all), molecular biology, equine, agronom	183	5
animal science and zoology, food science, veterinary	Alireza B.	Aghaie	food science, veterinary (all), food animals, zoon	14	0
animal science and zoology, food science, veterinary	M.	Khosravi Saberi	animal science and zoology, food science, veterina	0	0
animal science and zoology, food science, veterinary	Morteza	Bitaraf Sani	food animals, microbiology, veterinary (parasitology)	36	4
animal science and zoology, food science, veterinary	Seied	Esmaili-Nasani	small animals, animal science and zoology, food an	135	7
animal science and zoology, veterinary (all)	Hadi	wang	veterinary (all), animal science and zoology	2	1
animal science and zoology, veterinary (all)	Yingfan	Xu	ecological modeling, business and international ma	4	1
animal science and zoology, veterinary (all)	Yan	Yu	veterinary (all), control and systems engineering	13	2

Εικόνα 4.9: Τμήμα πίνακα συσχέτισης με τα metadata profiles συγγραφέων (πηγή: ιδία χρήση).

στοιχίζονται στην ίδια σειρά. Κατά αυτόν τον τρόπο μπορεί να διερευνηθεί εκ μέρους των χρηστών το κάθε ζεύγος και σε περίπτωση που διαπιστωθεί η ύπαρξη διπλοτύπων, μπορούν να επιλέξουν το αντίγραφο προς διαγραφή. Με το button ‘Clear Duplicates’ πραγματοποιείται η απαλοιφή διπλοτύπων ενώ μέσω του button ‘Go Back’ είναι δυνατή η επιστροφή στο κεντρικό menu διαχείρισης διπλότυπων.



Εικόνα 4.11: Menu διαχείρισης διπλοτύπων για το Σενάριο Αναζήτησης 1 (πηγή: ίδια χρήση).



Εικόνα 4.12: Επιλογή διπλοτύπων προς διαγραφή για το Σενάριο Αναζήτησης 1 (πηγή: ίδια χρήση).

Το τελικό βήμα της διαχείρισης των αποτελεσμάτων αναζήτησης είναι η εξαγωγή των metadata σε αρχείο Excel ή/και CSV, κάτι που δίνει την δυνατότητα στον χρήστη να αξιοποιήσει και άλλα εργαλεία για την στατιστική ανάλυση των βιβλιομετρικών δεδομένων που σύλλεξε μέσω της πλατφόρμας Intelnetus. Σημειώνεται πως οποιαδήποτε τροποποίηση εφαρμοστεί στα τελικά δεδομένα του πίνακα (π.χ εφαρμογή φίλτρου, διαγραφή καταχωρήσεων ως διπλότυπα) μεταφέρεται αυτόματα και στα αρχεία εξόδου. Στις Εικόνες 4.13 και 4.14 παρουσιάζονται παραδείγματα αρχείων εξόδου Excel και CSV αντίστοιχα.

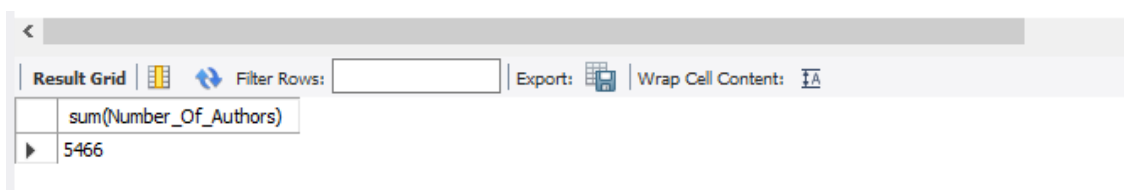
Εικόνα 4.13: Δείγμα αρχείου εξόδου Excel με τα τελικά metadata (πηγή: ιδία χρήση).

Εικόνα 4.14: Δείγμα αρχείου εξόδου CSV με τα τελικά metadata (πηγή: ιδία χρήση).

Σκοπός του συγκεκριμένου σεναρίου είναι η ανάκτηση ενός μεγαλύτερου πλήθους metadata δημοσιεύσεων, ώστε να αξιολογηθεί η αποτελεσματικότητα της πλατφόρμας στην ανάκτηση και την διαχείριση δεδομένων υπό πιο απαιτητικές συνθήκες λειτουργίας. Οι παράμετροι αναζήτησης που καθορίστηκαν είναι:

- 66

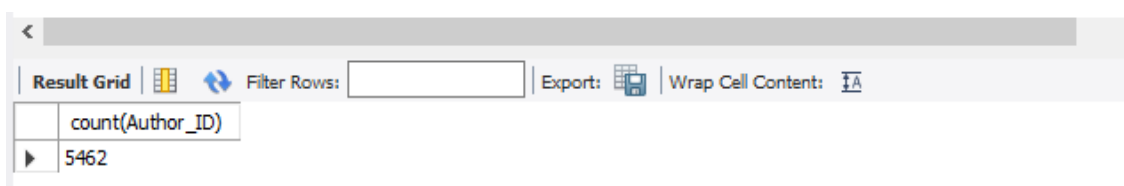
Το πλήθος των καταχωρήσεων που επεστράφησαν ανέρχεται στις 4012 ενώ το αναμενόμενο και το πραγματικό πλήθος διασυνδέσεων άρθρων – συγγραφέων και άρθρων – οργανισμών υπολογίζονται μέσω των queries στις Εικόνες 4.15, 4.16, 4.17 και 4.18. Το ποσοστό επιτυχών αποθηκεύσεων metadata συγγραφέων για το καθορισμένο dataset δημοσιεύσεων είναι 99.92%, ενώ το αντίστοιχο για τους οργανισμούς είναι 99.84%. Παρατηρείται λοιπόν μια μικρή υποχώρηση στην αποδοτικότητα του μηχανισμού εξαγωγής σε αυτό το σενάριο αναζήτησης σε σχέση με το Σενάριο 1 (0.08% για συγγραφείς και 0.16% για οργανισμούς), για μία αύξηση του dataset κατά 1000% (387 καταχωρήσεις για το Σενάριο 1 και 4012 καταχωρήσεις για το Σενάριο 2).



The screenshot shows a software interface with a toolbar at the top containing icons for 'Result Grid', 'Filter Rows', 'Export', and 'Wrap Cell Content'. Below the toolbar is a table with one row. The first column contains a right-pointing triangle icon, and the second column contains the text '5466'.

	sum(Number_Of_Authors)
▶	5466

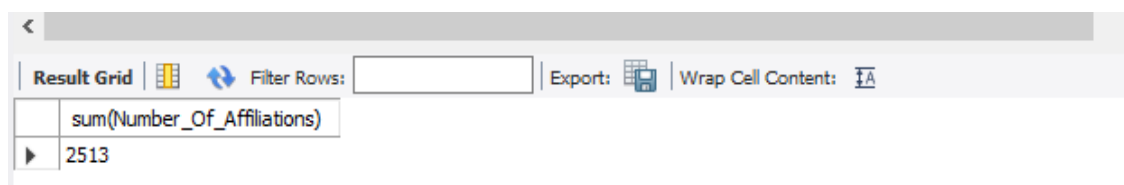
Εικόνα 4.15: Αναμενόμενο πλήθος διασυνδέσεων άρθρων – συγγραφέων (πηγή: ίδια χρήση).



The screenshot shows a software interface with a toolbar at the top containing icons for 'Result Grid', 'Filter Rows', 'Export', and 'Wrap Cell Content'. Below the toolbar is a table with one row. The first column contains a right-pointing triangle icon, and the second column contains the text '5462'.

	count(Author_ID)
▶	5462

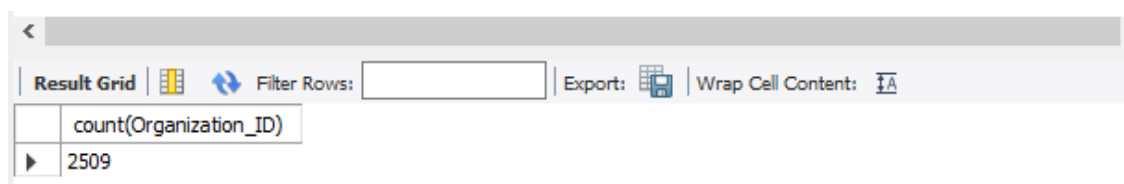
Εικόνα 4.16: Πραγματικό πλήθος διασυνδέσεων άρθρων – συγγραφέων (πηγή: ίδια χρήση).



The screenshot shows a software interface with a toolbar at the top containing icons for 'Result Grid', 'Filter Rows', 'Export', and 'Wrap Cell Content'. Below the toolbar is a table with one row. The first column contains a right-pointing triangle icon, and the second column contains the text '2513'.

	sum(Number_Of_Affiliations)
▶	2513

Εικόνα 4.17: Αναμενόμενο πλήθος διασυνδέσεων άρθρων – οργανισμών (πηγή: ίδια χρήση).

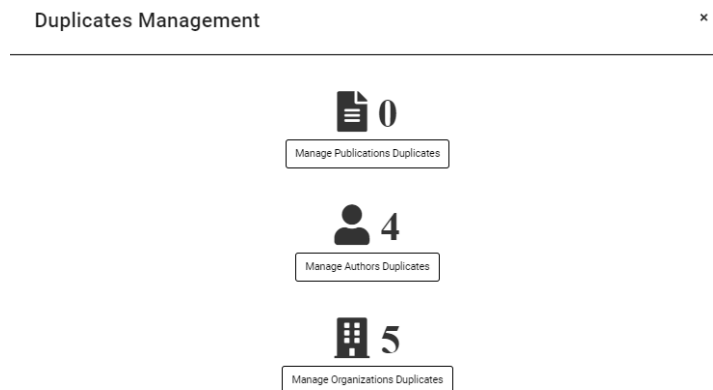


The screenshot shows a software interface with a toolbar at the top containing icons for 'Result Grid', 'Filter Rows', 'Export', and 'Wrap Cell Content'. Below the toolbar is a table with one row. The first column contains a right-pointing triangle icon, and the second column contains the text '2509'.

	count(Organization_ID)
▶	2509

Εικόνα 4.18: Πραγματικό πλήθος διασυνδέσεων άρθρων – οργανισμών (πηγή: ίδια χρήση).

Στο menu διαχείρισης διπλοτύπων της Εικόνας 4.19 παρατηρείται ότι 4 ζεύγη profiles συγγραφέων και 5 ζεύγη οργανισμών επισημαίνονται ως πιθανά διπλότυπα, ενώ δεν υπάρχουν αντίστοιχες επισημάνσεις που να αφορούν το σύνολο καταχωρήσεων δημοσιεύσεων. Οι καταγραφές προς διερεύνηση για συγγραφείς και οργανισμούς παρουσιάζονται στους πίνακες των Εικόνων 4.20 και 4.21 αντίστοιχα.



Εικόνα 4.19: Menu διαχείρισης διπλοτύπων για το Σενάριο Αναζήτησης 2 (πηγή: ιδία χρήση).

The image shows a 'Duplicates Management' window with two tables side-by-side: 'Primary Variants' and 'Secondary Variants'. Both tables have columns for Name, Citations, Index, and a checkbox.

Primary Variants				Secondary Variants			
Name	Citations	Index		Name	Citations	Index	
Shuxin Zhou	2	1	☐	Mingliang Zhou	243	11	☐
Mid Zubair Hasan	0	0	☐	Mid Mehedi Hasan	233	8	☐
Yingjie Tian	54	3	☐	Yingjie Tian	5091	32	☐
Hao Wu	109	7	☐	Jialun Wu	36	4	☐

At the bottom right, there are buttons for 'Go Back' and 'Clear Duplicates'.

Εικόνα 4.20: Πιθανά διπλότυπα συγγραφέων για το Σενάριο Αναζήτησης 2 (πηγή: ιδία χρήση).

The image shows a 'Duplicates Management' window with two tables side-by-side: 'Primary Variants' and 'Secondary Variants'. Both tables have columns for Name and a checkbox.

Primary Variants		Secondary Variants	
Name		Name	
Indian Institute of Technology Kanpur	☐	Indian Institute of Technology Kharagpur	☐
Trinity Earth Technology Co.Ltd	☐	Trinity Earth Technology Co.Ltd	☐
Chongqing Meliansen Technology Co. Ltd	☐	Chongqing Meliansen Technology Co. Ltd.	☐
Rajshahi University	☐	Rajshahi University of Engineering and Technology	☐
Changzhou College of Information Technology	☐	Changzhou College of Information Technology	☐

At the bottom right, there are buttons for 'Go Back' and 'Clear Duplicates'.

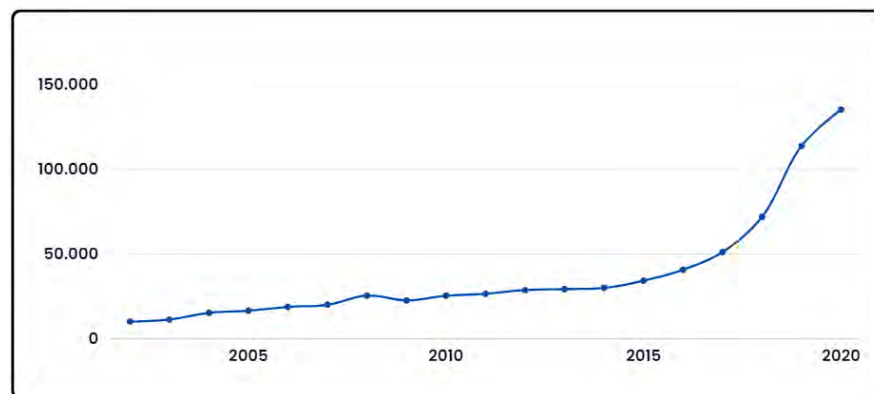
Εικόνα 4.21: Πιθανά διπλότυπα συγγραφέων για το Σενάριο Αναζήτησης 2 (πηγή: ιδία χρήση).

4.3.3 Σενάριο αναζήτησης 3 – Εκτίμηση πλήθους δημοσιεύσεων

Στόχος του συγκεκριμένου σεναρίου είναι η εκτίμηση του πλήθους δημοσιεύσεων των οποίων τα metadata μπορούν να ανακτηθούν μελλοντικά από την πλατφόρμα, αφού έχει εξασφαλιστεί η απαραίτητη υλικοτεχνική υποδομή για την καταχώρηση μεγάλου όγκου δεδομένων. Οι παράμετροι εισόδου αυτού του σεναρίου είναι:

- ‘artificial intelligence’, ‘machine learning’, ‘deep learning’, ‘learning algorithm’ και ‘pattern recognition’ ως λέξεις – κλειδιά και τελεστές OR.
- Χρονικό εύρος από το έτος 2002 μέχρι το 2020.
- ‘Computer Science’ και ‘Engineering’ ως επιστημονικά πεδία.

Αποτέλεσμα αυτού του σεναρίου αναζήτησης είναι ο εντοπισμός 727,058 δημοσιεύσεων που πληρούν τα παραπάνω κριτήρια. Η διακύμανση του αριθμού των δημοσιευμένων ερευνών ανά τα έτη αποτυπώνεται στο Διάγραμμα 4.1 παρακάτω. Η αναλυτική παράθεση του αριθμού των δημοσιεύσεων για κάθε έτος ξεχωριστά προβάλλεται στα logs του τερματικού της Εικόνας 4.22 που δημιουργήθηκαν κατά την εκτέλεση του κώδικα.



Διάγραμμα 4.1 Αριθμός δημοσιεύσεων ανά έτος για το Σενάριο Αναζήτησης 3 (πηγή: ιδία χρήση).

```
10180 publications were detected for year 2002.  
11349 publications were detected for year 2003.  
15325 publications were detected for year 2004.  
16577 publications were detected for year 2005.  
18816 publications were detected for year 2006.  
20091 publications were detected for year 2007.  
25474 publications were detected for year 2008.  
22661 publications were detected for year 2009.  
25390 publications were detected for year 2010.  
26511 publications were detected for year 2011.  
28728 publications were detected for year 2012.  
29285 publications were detected for year 2013.  
30016 publications were detected for year 2014.  
34262 publications were detected for year 2015.  
40673 publications were detected for year 2016.  
51073 publications were detected for year 2017.  
71850 publications were detected for year 2018.  
113703 publications were detected for year 2019.  
135094 publications were detected for year 2020.
```

Εικόνα 4.22: Αναλυτική παράθεση αριθμού δημοσιεύσεων ανά έτος για το Σενάριο Αναζήτησης 3 (πηγή: ιδία χρήση).

4.4 Σύνοψη κεφαλαίου

Σε αυτό το κεφάλαιο πραγματοποιήθηκε η παρουσίαση των αποτελεσμάτων της διπλωματικής εργασίας. Αρχικά έγινε περιγραφή του UI, το οποίο οργανώνεται σε δύο σελίδες: την Αρχική Σελίδα και την Σελίδα Αναζήτησης. Στην Αρχική Σελίδα γίνεται μια συνοπτική περιγραφή των δυνατοτήτων και των μελλοντικών επεκτάσεων της πλατφόρμας Intelnetus που αναπτύσσεται στην παρούσα εργασία. Στην Σελίδα Αναζήτησης περιλαμβάνεται μια φόρμα μέσω της οποίας ο χρήστης μπορεί να εισάγει τις παραμέτρους αναζήτησης που επιθυμεί, προκειμένου να ξεκινήσει η διαδικασία της άντλησης metadata από την πλατφόρμα της Scopus και την οργάνωσή τους σε μια ενιαία σχεσιακή δομή πίνακα.

Για την εκτίμηση της αποτελεσματικότητας τόσο του μηχανισμού εξαγωγής δεδομένων όσο και των λειτουργιών διαχείρισης των τελικών αποτελεσμάτων πραγματοποιήθηκαν τρία σενάρια αναζήτησης:

1. Σενάριο Αναζήτησης 1: Σκοπός αυτού του σεναρίου είναι μια αρχική επαλήθευση της αποτελεσματικότητας και της ομαλής λειτουργίας της πλατφόρμας κατά την διαχείριση ενός περιορισμένου πλήθους metadata, ώστε να είναι εφικτή η παρουσίαση σε ένα εύλογο χρονικό διάστημα. Διαπιστώθηκε πως ο μηχανισμός εξαγωγής και καταχώρησης δεδομένων στη βάση λειτουργεί στο ακέραιο με μηδαμινή απώλεια δεδομένων, ενώ και οι λειτουργίες διαχείρισης των αποτελεσμάτων, όπως ταξινόμηση και φιλτράρισμα των αποτελεσμάτων, η διαχείριση πιθανών διπλοτύπων και η εξαγωγή δεδομένων πραγματοποιούνται όπως αναμενόταν.
2. Σενάριο Αναζήτησης 2: Σε αυτό το σενάριο τέθηκαν οι ίδιοι στόχοι με αυτούς στο Σενάριο 1 αλλά για μεγαλύτερο όγκο δεδομένων και για εκτενέστερο χρονικό διάστημα. Η μόνη διαφορά που διαπιστώθηκε ήταν μια μικρή ελάττωσή της αποτελεσματικότητας της διαδικασίας εξαγωγής metadata συγγραφέων και οργανισμών για το dataset δημοσιεύσεων που ανακτήθηκε.
3. Σενάριο Αναζήτησης 3: Η διαδικασία αναζήτησης αυτού του σεναρίου αποσκοπεί αποκλειστικά στην εκτίμηση του πλήθους των δημοσιεύσεων των οποίων τα metadata μπορούν να ανακτηθούν μελλοντικά με την κατάλληλη υλικοτεχνική υποστήριξη.

Οι τεχνικές λεπτομέρειες σχετικά με τα σενάρια αναζήτησης που αναφέρθηκαν παραθέτονται στον Πίνακα 4.1 παρακάτω:

Πίνακας 4.1 Τεχνικές λεπτομέρειες σεναρίων αναζήτησης (πηγή: ίδια χρήση)

Σενάριο Αναζήτησης	Λέξεις - Κλειδιά	Χρονικό Εύρος	Επιστημονικά Πεδία	Πλήθος δεδομένων που επεστράφησαν ή εντοπίστηκαν	Ποσοστό επιτυχών καταχωρήσεων συγγραφεών	Ποσοστό επιτυχών καταχωρήσεων οργανισμών
1	Artificial intelligence, Machine learning, Deep learning, Learning Algorithm, Pattern Recognition	2022	Social Sciences	387	100%	100%
2	Ομοίως με 1	2020 - 2022	Computer Science, Engineering	4012	99.92%	99.84%
3	Ομοίως με 1	2002 - 2020	Ομοίως με 2	727.058	-	-

Κεφάλαιο 5 ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1 Σύνοψη και συμπεράσματα

Το κύριο ερώτημα που τέθηκε στο πλαίσιο της παρούσας διπλωματικής εργασίας είναι το πως μπορούν οι έρευνες που μελετούν τα μοτίβα συνεργασιών στα δίκτυα γνώσης να γίνονται πιο σύντομα και με μεγαλύτερη ευκολία. Έτσι, ως σκοπός της διπλωματικής εργασίας ορίστηκε η ανάπτυξη μιας διαδικτυακής εφαρμογής η οποία θα μπορεί να αξιοποιηθεί από ερευνητές για την αυτοματοποίηση της διαδικασίας άντλησης και οργάνωσης μεταδεδωμένων επιστημονικής βιβλιογραφίας. Έπειτα από διερεύνηση πολλαπλών εργαλείων επεξεργασίας και στατιστικής ανάλυσης βιβλιομετρικών δεδομένων που έχουν ήδη αναπτυχθεί, διαπιστώθηκε ότι κανένα εξ αυτών δεν παρέχει τη δυνατότητα αυτόματης άντλησης δεδομένων και οργάνωση τους σε μια ενιαία σχεσιακή δομή πίνακα, ώστε να είναι δυνατή σε σύντομο χρονικό διάστημα η στατιστική τους ανάλυση.

Το κενό αυτό στην υπάρχουσα βιβλιογραφία καλείται να καλύψει η διπλωματική εργασία, με την ανάπτυξη της εφαρμογής Intelnetus. Η πλατφόρμα επιτρέπει στους χρήστες της να εισάγουν τα κριτήρια αναζήτησης που επιθυμούν, προκειμένου να συλλέξουν τα μεταδεδωμένα των επιστημονικών δημοσιεύσεων που πληρούν αυτά τα κριτήρια. Παράλληλα, γίνεται και ο αυτόματος συσχετισμός των δημοσιεύσεων με τους ερευνητές και τους αντίστοιχους οργανισμούς που συμμετείχαν στην συγγραφή τους. Έπειτα από την διεξαγωγή τριών σεναρίων αναζήτησης παρατηρήθηκε ότι ο μηχανισμός εξαγωγής και διαχείρισης δεδομένων της πλατφόρμας παρέχει υψηλή απόδοση με την ακεραιότητα της πληροφορίας που φτάνει στον χρήστη να προσεγγίζει το 100%, ενώ οι λειτουργίες δυναμικής διαχείρισης των αποτελεσμάτων όπως η εξαγωγή δεδομένων από την πλατφόρμα μέσω αρχείων Excel και CSV και η επιλογή πιθανών διπλοεγγραφών προς διαγραφή εκ μέρους των χρηστών τελούνται αποτελεσματικά.

Η μεγαλύτερη πρόκληση που αντιμετωπίστηκε κατά την ανάπτυξη της πλατφόρμας είναι η ομαλή διαχείριση και καταχώρηση στη βάση SQL των αντλούμενων δεδομένων. Συγκεκριμένα, παρουσιάστηκαν περιπτώσεις αντικειμένων των οποίων ορισμένα χαρακτηριστικά, είτε είχαν τιμή null είτε απουσίαζαν εντελώς από το σώμα αυτών των αντικειμένων. Η ανάγνωση αυτών των χαρακτηριστικών οδηγούσε συχνά σε σφάλματα που είτε απέτρεπαν την καταχώρηση δεδομένων στη βάση είτε ακόμα τερμάτιζαν την διαδικασία άντλησης. Ο τρόπος που αντιμετωπίστηκε η συγκεκριμένη δυσκολία είναι η υλοποίηση

μεθόδων ελέγχου. Οι μέθοδοι αυτές είναι υπεύθυνες για την αντικατάσταση null τιμών (σε περίπτωση που υπάρχουν εντός ενός αντικειμένου) με τιμές που μπορούν να διαχειριστούν αποτελεσματικά από το πρόγραμμα, ώστε να αποφευχθεί η δημιουργία σφαλμάτων και η απώλεια δεδομένων.

Κατά αυτόν τον τρόπο προσφέρεται μια απλοποιημένη και πιο αποδοτική εμπειρία άντλησης και αξιοποίησης metadata επιστημονικών δημοσιεύσεων στους ερευνητές για την αποπεράτωση εμπειρικών ερευνών με μεγαλύτερη ευκολία και ταχύτητα.

5.2 Μελλοντικές επεκτάσεις

Οι δυνατότητες της πλατφόρμας Intelnetus στο πλαίσιο αυτής της διπλωματικής εργασίας, αποτελούν τη βάση για την περαιτέρω εξέλιξη της πλατφόρμας με την ενσωμάτωση νέων δυνατοτήτων στο λειτουργικό εύρος της. Παρακάτω παραθέτονται προτάσεις σχετικά με το ποιες μπορεί να είναι οι μελλοντικές επεκτάσεις της παρούσας έρευνας, αλλά και το ποιες είναι οι απαραίτητες προϋποθέσεις που πρέπει να καλυφθούν για την πραγμάτωση αυτών.

Στην περαιτέρω εξέλιξη της πλατφόρμας Intelnetus και της αναβάθμισης του ρόλου της στην εκπόνηση επιστημονικομετρικών μελετών, θα μπορούσε να συνεισφέρει:

- Η πρόσβαση στο περιεχόμενο περισσότερων βάσεων δεδομένων επιστημονικών δημοσιεύσεων μέσω των αντίστοιχων APIs, για τον εμπλουτισμό της βιβλιομετρικής πληροφορίας που παρέχεται στους χρήστες αλλά και για την διασταύρωση εγκυρότητας των αντλούμενων δεδομένων.
- Εγκατάσταση της πλατφόρμας και της βάσης της σε κατάλληλο εξυπηρετητή (server), ώστε να μην επηρεάζεται αρνητικά η απόδοση της σε περιπτώσεις άντλησης, διαχείρισης και αποθήκευσης μεγάλου όγκου δεδομένων.
- Ανάπτυξη αλγορίθμων για την μέτρηση των διάφορων διαστάσεων της εγγύτητας μεταξύ των κόμβων του εκάστοτε δικτύου γνώσης και τον υπολογισμό της στατιστικής σημασίας της κάθε εγγύτητας.
- Εξαγωγή σε ξεχωριστό αρχείο εξόδου των μετρήσεων εγγύτητας.
- Διερεύνηση δυνατότητας απεικόνισης δικτύων γνώσης και δημιουργία στατιστικών γραφημάτων – πινάκων.
- Ανάπτυξη AI μοντέλου γλώσσας για την αποτελεσματικότερη κατηγοριοποίηση των οργανισμών και για τον αποδοτικότερο εντοπισμό διπλοεγγραφών.

Λαμβάνοντας υπόψιν τις παραπάνω προοπτικές, γίνεται αντιληπτό πως η πλατφόρμα Intelnetus έχει τη δυναμική να εξελιχθεί σε ένα πρωτοπόρο και καινοτόμο εργαλείο που θα συνεισφέρει σημαντικά στην προσπάθεια μελέτης και κατανόησης δικτύων γνώσης.

Βιβλιογραφία

- [1] Garcia-Fuentes, W. και συν., 2022. A 10-Year Scientometrics Analysis of Brain Tumors Treated with Gamma Knife Radiosurgery: Visualization, Characteristics, and Scientific Trends. *Hindawi*, 10 3.
- [2] Zeng, A. και συν., 2017. The science of science: From the perspective of complex systems. *Physics Reports*, 27 10, pp. 1-73.
- [3] Bordons, M., Javier, A., Gonzalez-Albo, B. & Diaz-Faes, A., 2014. The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informatics*, 26 12, pp. 135-144.
- [4] Newman, M., 2004. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Proceedings of the national academy of sciences*, pp. 5200-5205.
- [5] Newman, M. E. J., 2001. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physica; Review*, 28 6.
- [6] Kumar, S., 2015. Co-authorship networks: a review of the literature. *Aslib Journal of Information Management*.
- [7] Melin, G. & Persson, O., 1996. Studying research collaboration using co-authorships. *Scientometrics*, pp. 363-377.

- [8] Barnet, A., Ault, R. & Kaserman, D., 1988. The rising incidence of co-authorship in economics: Further evidence. *The Review of Economics and Statistics*, pp. 539-543.
- [9] Guan, J., Yan, Y. & Zhang, J., 2017. The impact of collaboration and knowledge networks on citations. *Journal of Infometrics*, pp. 407-422.
- [10] Phelps, C., Heidl, R. & Wadhwa, A., 2012. Knowledge, networks and knowledge networks: A review and research agenda. *Journal of Management*, pp. 1115-1166.
- [11] Pauli, J., Basso, K., Gobi, R. L. & Bilhar, A., 2020. The effect of co-authorship network density on the performance of postgraduate programs. *Brazilian Business Review*, pp. 576-588.
- [12] Rossoni, L. & Hocauen-da-Silva, A. J., 2008. Cooperacao entre pesquisadores da area de administracao da informacao: evidencias estruturais de fragmentacao das relacoes no campo científico. *Revista de administracao*, pp. 138-151.
- [13] Bernela, B. & Mirald, B., 2016. Studuing co-authorship network dynamics throughout geographical trajectories of researchers: what part does mobility play?.
- [14] Broekel, T., 2015. The co-evaluation of proximities - A network level study. *Regional studies*, pp. 921-935.
- [15] Boschma, R., 2005. Role of proximity in interaction and performance: Conceptual and empirical challenges.
- [16] Burt, R. S., 2000. The network structure of social capital. *Rsearch in organizational behavior*, pp. 345-323.
- [17] Huggins, R., Johnston, A. & Thompson, P., 2012. Network capital, social capital, and knowledge flow: how the nature of inter-organizational networks impacts on innovation. *Industry and Innovation*, pp. 203-232.

- [18] Inkpen, A. C. & W., T. E., 2005. Social capital, networks and knowledge transfer. *Academy of management review*, pp. 146-165.
- [19] Burt, R. S., 2004. From structural holes: The social structure of competition. *The new economic sociology: a reader*, pp. 325-348.
- [20] Granovetter, M. S., 1973. The strength of weak ties. *Americal journal of sociology*, pp. 407-422.
- [21] Khurana, P., Ganesan, G., Kumar, G. & Sharma, K., 2022. A Comparative Analysis of Unified Informetrics with Scopus and Web of Science. *Journal of Scientometric Res*, 8 7, pp. 146-154.
- [22] Khurana, P., Parul, K., Sharma, K. & Khatter, K., 2022. Proof of bibliometric indicators: a blockchain based consensus protocol for publications. *Multimedia Tools and Applications*, 18 11.
- [23] Aria, C., 2017. bibliometrix .. An R-tool for comprehensive science mapping. *Journal of Informetrics*, 12 9, pp. 959-975.
- [24] Dervis, H., 2019. Bibliometric Analysis using Bibliometrix an R Package. *Journal of Scientometric Res.*, 11 2.
- [25] Jan van Eck, N. & Waltman, L., 2009. Software survey: VOSviewer, a computer program. *Scientometrics*, 31 12, pp. 523-538.
- [26] Cobo, M., Lopze-Herrera, A., Herrera-Viedma, E. & Herrera, F., 2011. Science Mapping Software Tools: Review, Analysis, and Cooperative Study Among Tools. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 10 2, pp. 1382-1402.

- [27] Sangam, S. L. & Mogali, S. S., 2012. Mapping and Visualization Softwares tools: a review. *Conference Paper*, 2.
- [28] Jan van Eck, N. & Waltman, L., 2014. CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, 27 8, pp. 802-823.
- [29] Cobo, M., Lopze-Herrera, A., Herrera-Viedma, E. & Herrera, F., 2012. SciMAT: A New Science Mapping Analysis Software Tool. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 28 2, pp. 1609-1630.
- [30] Solomon, D., 2015. A different view: An inquiry into visualization of bibliometric data. *American Society for Engineering Education*, 14 6.
- [31] M., L. D. & Alpi, K. M., 2017. Bibliometric Network Analysis and Visualization for Serials Librarians: An Introduction to Sci2. *Serials Review*, 04 10, pp. 239-245.
- [32] Yang, J., Cheng, C., Shen, S. & Yang, S., 2017. Comparison of Complex Network Analysis Software: Citespace, SCI2 and Gephi. *IEEE 2nd International Conference on Big Data Analysis*.
- [33] Chen, C., 2005. CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 7 2, pp. 359-377.
- [34] Fredrik, A., Rickard, D., Birger, L. & Jesper, S., 2009. How to use Bibexcel for various types of bibliometric. *Celebrating Scholarly Communication Studies*, 5 6.
- [35] Pradhan, P., 2016. Science Mapping and Visualization Tools used in Bibliometric & Scientometric Studies: An Overview. *INFLIBNET*, 10.

- [36] Ruiz-Rosero, J., Ramirez-Gonzalez, G. & Viveros-Delgado, J., 2019. Software survey: ScientoPy, a scientometric tool for topics trend analysis in scientific publications. *Scientometrics*, 30 8, pp. 1165-1188.
- [37] Cestero, J. et al., 2022. Pysurveillance: A Novel Tool for Supporting Researchers in the Systematic Literature Review Process.
- [38] Παπαδοπούλου, Ε. & Βασιλακόπουλος, Μ., 2017. *Development of a web tool for database modeling*. Βόλος: s.n.
- [39] Rose, M. E. & Kitchin, J. R., 2019. pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. *SoftwareX*, 19 6.
- [40] Lammey, R., 2019. The changing landscape of scholarly identifiers: New opportunities with CrossRef's support for grants, preprints, and protocols. *Learned Publishing*, pp. 221-225.
- [41] Habibzadeh, M., Ghandehari, M. & Ghasemi, A., 2015. Reliability of DOI References in Biomedical Articles. *Archives of Iranian Medicine*, pp. 686-689.
- [42] Fenner, M., Lin, J. & Hepp, M., 2012. An empirical analysis of CrossRef, DataCite, and Dryad.. *D-Lib Magazine*.

ΠΑΡΑΡΤΗΜΑΤΑ

Παράρτημα Α

Ευρετήριο Ορολογίας

API: Η διεπαφή επικοινωνίας και αλληλοεπίδρασης μεταξύ δύο ή περισσότερων λογισμικών.

API Wrapper: Πακέτο ή βιβλιοθήκη γλώσσας προγραμματισμού που ενσωματώνει ένα ή περισσότερα APIs στις κλάσεις, τις συναρτήσεις και τις μεθόδους που περιέχει.

Blockchain: Κρυπτογραφημένο και αποκεντρωμένο σύστημα κατανεμημένου καταλόγου (ledger) για την καταγραφή και την επαλήθευση συναλλαγών μεταξύ διαφόρων συμμετεχόντων.

Boolean Τελεστής: Δυναδικός τελεστής (and, or, not).

Code-free UI: Διεπαφή χρήστη που δεν προϋποθέτει την γνώση ή χρήση κώδικα προγραμματισμού για την τέλεση των λειτουργιών που την χαρακτηρίζουν.

Dataframes: Δομές δεδομένων που οργανώνουν τα δεδομένα σε μορφή πίνακα με σειρές και στήλες, παρέχοντας ευκολία στην ανάλυση και την επεξεργασία τους.

DLT Εφαρμογή: Εφαρμογή που χρησιμοποιεί τεχνολογία κατανεμημένου καταλόγου για την πραγματοποίηση και την επικύρωση συναλλαγών και την ανταλλαγή πληροφοριών.

H-Index: Βιβλιομετρική ένδειξη που αντιστοιχεί στο πλήθος των δημοσιεύσεων ενός συγγραφέα που έχουν λάβει πάνω από έναν συγκεκριμένο αριθμό παραπομπών από άλλες δημοσιεύσεις.

HTTP Αίτημα: Αίτημα που αποστέλλεται από μία μηχανή – πελάτη σε μία μηχανή – εξυπηρετητή με σκοπό την ανταλλαγή δεδομένων και την μετάδοση πληροφορίας.

RDBMS: Σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων που αξιοποιείται για τη δημιουργία και διαχείριση σχημάτων βάσεων δεδομένων, καθώς και για την μεταγλώττιση και την εκτέλεση SQL queries.

Scraper Script: Αρχείο κώδικα προγραμματισμού που διατρέχει αυτόματα διαδικτυακές σελίδες και ιστότοπους για την εξαγωγή δεδομένων.

UI: Η διεπαφή μιας εφαρμογής ή ενός προγράμματος μέσω της οποίας οι χρήστες έχουν τη δυνατότητα αξιοποίησης των λειτουργιών του εκάστοτε λογισμικού.

Δίκτυα Γνώσης: Δομημένες αναπαραστάσεις δικτύων που αποτυπώνουν τις συνδέσεις και τις και τις συσχετίσεις μεταξύ οντοτήτων και θεσμών εντός μιας επιστημονικής κοινότητας.

Μεταδεδομένα: Δομημένα σύνολα δεδομένων που αποσκοπούν στην παροχή χαρακτηριστικών και περιγραφής άλλων δεδομένων για την καλύτερη κατανόηση και διαχείριση τους.

Πίνακες Συσχέτισης: Οργανωμένες δομές δεδομένων που αποτελούνται από σειρές και στήλες με σκοπό την καταχώρηση πληροφορίας που εκφράζει την σύνδεση μεταξύ πολλαπλών οντοτήτων υπό ένα συγκεκριμένο πλαίσιο.

Παράρτημα Β

Κατάλογος φακέλων και αρχείων στο Git Hub repository

- intelnetus-api/Scopus (το API της πλατφόρμας)
 - ConnectToMySQL (σύνδεση με την SQL βάση)
 - Connector.py
 - DuplicatesDetection (μέθοδοι εντοπισμού διπλοτύπων)
 - AuthorsDuplicates.py
 - OrganizationsDuplicates.py
 - PublicationsDuplicates.py
 - Entities (ορισμός κλάσεων δημοσιεύσεων, συγγραφέων και οργανισμών)
 - Classes.py
 - Fuzzywuzzy (εξωτερική βιβλιοθήκη για την σύγκριση strings)
 - MetadataManipulation (εξαγωγή και διαχείριση metadata)
 - Extractor.py
 - Preprocessing (προ-επεξεργασία metadata)
 - Methods.py
 - WebScraper (αποστολή HTTP αιτημάτων για την λήψη δεδομένων)
 - Methods.py
 - IntelnetusAPI.py (έναρξη λειτουργίας και αρχικοποίηση API)
- intelnetus-db
 - DatabaseSchema.txt (SQL queries για τη δημιουργία του σχήματος βάσης)
- intelnetus-ui/angular/scr/app
 - home (component για την αρχική σελίδα)
 - search (component για την σελίδα αναζήτησης)

Για μετάβαση στο Git Hub repository πατήστε [εδώ](#).