Ενοποιημένη διαχείριση πολλαπλών βάσεων αναφορών (Citation Indexes) για τον εμπλουτισμό Ιδρυματικών Αποθετηρίων

Δημήτριος Κουής[1], Γιώργος Βεράνης[1], Μάριος Ζέρβας[2], Πέτρος Αρτέμης[2], Ανδρέας Γιαννακόπουλος[1], Χρήστος Μπέλλας[3]

[1] Τμήμα Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης, Πανεπιστήμιο Δυτικής Αττικής, Αιγάλεω, Αθήνα, dkouis@uniwa.gr / gveranis@gmail.com / gianandr4@gmail.com

[2] Βιβλιοθήκη και Υπηρεσία Πληροφόρησης, Τεχνολογικό Πανεπιστήμιο Κύπρου, marios.zervas@cut.ac.cy / petros.artemi@cut.ac.cy

[2] Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, chribell@csd.auth.gr

ΠΑΝΕΛΛΗΝΙΟ ΦΑΡΟΙ ΣΥΝΕΔΡΙΟ ΕΛΕΥΘΕΡΙΑΣ ΑΚΑΔΗΜΑΪΚΩΝ ΤΗΣ ΒΙΒΛΙΟΘΗΚΩΝ ΓΝΩΣΗΣ

Department of Archival, Library and Information Studies

Information Management Research Lab

Εισαγωγή – Πλαίσιο – Επιδιωκόμενα αποτελέσματα

Ένα **ευρετήριο ή βάση αναφορών (citation index)** αποτελεί κάτι περισσότερο από μια απλή πηγή βιβλιογραφικών πληροφοριών, τα οποία είναι διασυνδεδεμένα μεταξύ τους, αφού παρέχει επιπλέον μια **αυστηρή δόμηση και ένα σαφώς καθορισμένο μοντέλο δεδομένων** (McVeigh, 2017).

Σήμερα υπάρχουν **πολλά ευρετήρια αναφορών** [Web of Science (Clarivate Analytics), Scopus (Elsevier), Google Scholar (Google), Microsoft Academic (Microsoft), Dimension (Digital Science & Research Solutions Inc.)], καθώς και **μεμονωμένες, εξειδικευμένες βάσεις** π.χ. PubMed. Το τοπίο συμπληρώνεται από μια σειρά από **υπηρεσίες** όπως π.χ. η ORCID, το ResearchGate κ.λπ., καθώς και από **παρόχους μοναδικών αναγνωριστικών** (PID - Persistent Identifiers) για ψηφιακά αντικείμενα όπως η CrossRef και η DataCite, οι οποίες αναπτύσσουν και **διατηρούν "γράφους" (graphs) από βιβλιογραφικά δεδομένα**.

Εισαγωγή – Πλαίσιο – Επιδιωκόμενα αποτελέσματα

Από τα πρώτα κιόλας χρόνια εμφάνισης των ευρετηρίων αναφορών, πίσω στην δεκαετία του 2000, γεννήθηκαν μια σειρά από ζητήματα όπως:

- **ποιο είναι το καταλληλότερο** μέσο για την αναζήτηση δεδομένων ανά επιστημονικό πεδίο (thematic coverage),
- **ποιο παρέχει την μεγαλύτερη κάλυψη** (number of sources indexed),
- **ποιο παρέχει τα πιο ακριβή και σωστά δεδομένα**,
- ποιο **παρέχει τους ακριβέστερους βιβλιομετρικούς δείκτες** κ.λπ.

Εισαγωγή – Πλαίσιο – Επιδιωκόμενα αποτελέσματα

Σε αυτό το πλαίσιο, διαφαίνεται ότι υπάρχει ανάγκη για να υπάρξει ένα εργαλείο που θα μπορούσε να παρέχει μια **ενοποιημένη διαχείριση των βιβλιογραφικών δεδομένων** που παρέχει το κάθε **ευρετήριο** αναφορών με έμφαση **στην αποδιπλοποίηση (deduplication) των κοινών δημοσιευμάτων**.

Το εργαλείο αυτό αφενός θα βοηθήσει στον **εμπλουτισμό των Ιδρυματικών Αποθετηρίων (αφορμή)** με έμφαση στην εξοικονόμηση ανθρωπίνων πόρων, αλλά μπορεί να συμβάλλει **στις διαδικασίες αναβάθμισης των ιστοσελίδων** των Ιδρυμάτων και των **διαδικασιών αξιολόγησης**.

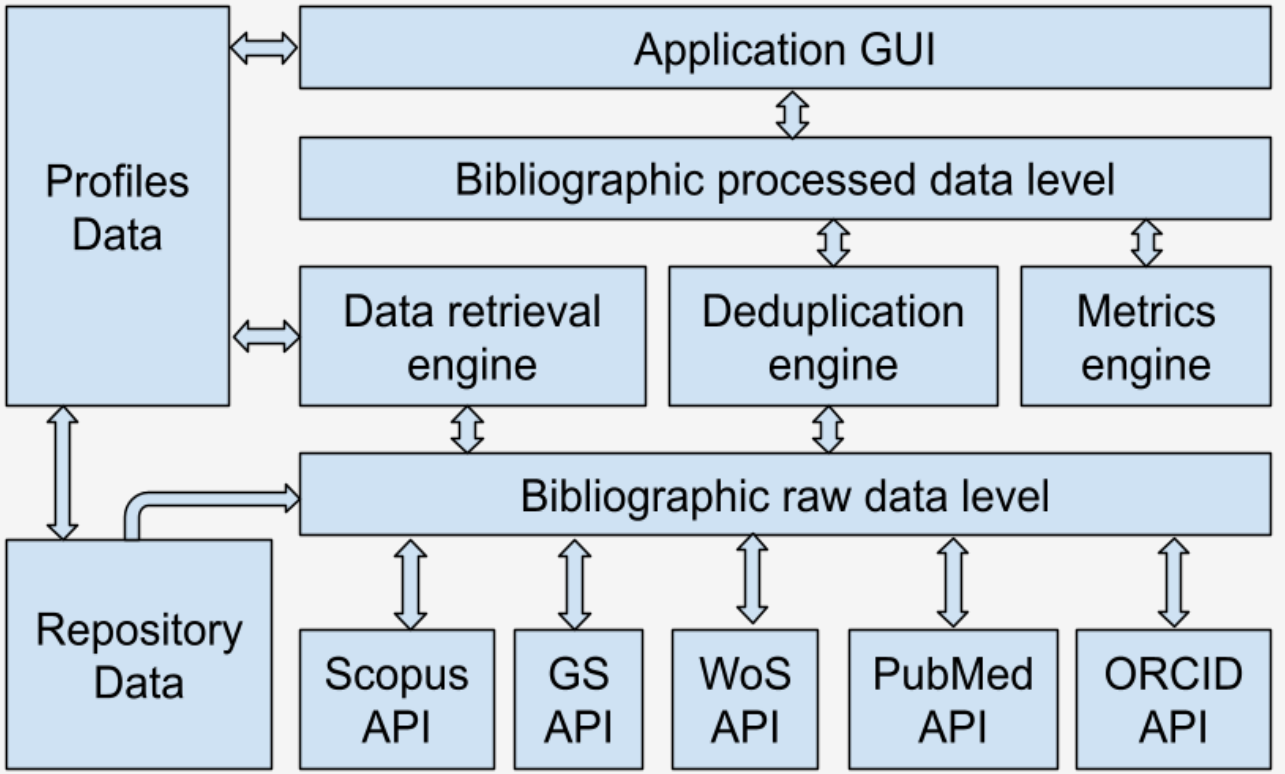
Μεθοδολογία - Υλοποίηση

Μελέτη και παραγωγή προδιαγραφών περιβάλλοντος

1. Επιλογή των βάσεων / ευρετηρίων αναφορών που θα συμμετέχουν στην άντληση των βιβλιογραφικών δεδομένων
2. Μελέτη των διαθέσιμων APIs και των δεδομένων που προσφέρει η κάθε βάση / ευρετηρίου αναφορών
3. Δημιουργία ενός ελάχιστου επιπέδου κοινών δεδομένων - Μοντέλο δεδομένων
4. Καθορισμός αλγόριθμου αποδιπλοποίησης εγγραφών από διαφορετικά ευρετήρια αναφορών - βιβλιομετρικοί δείκτες
5. Διεπαφή χρήση - Προσφερόμενες λειτουργίες - Δεδομένα εξόδου

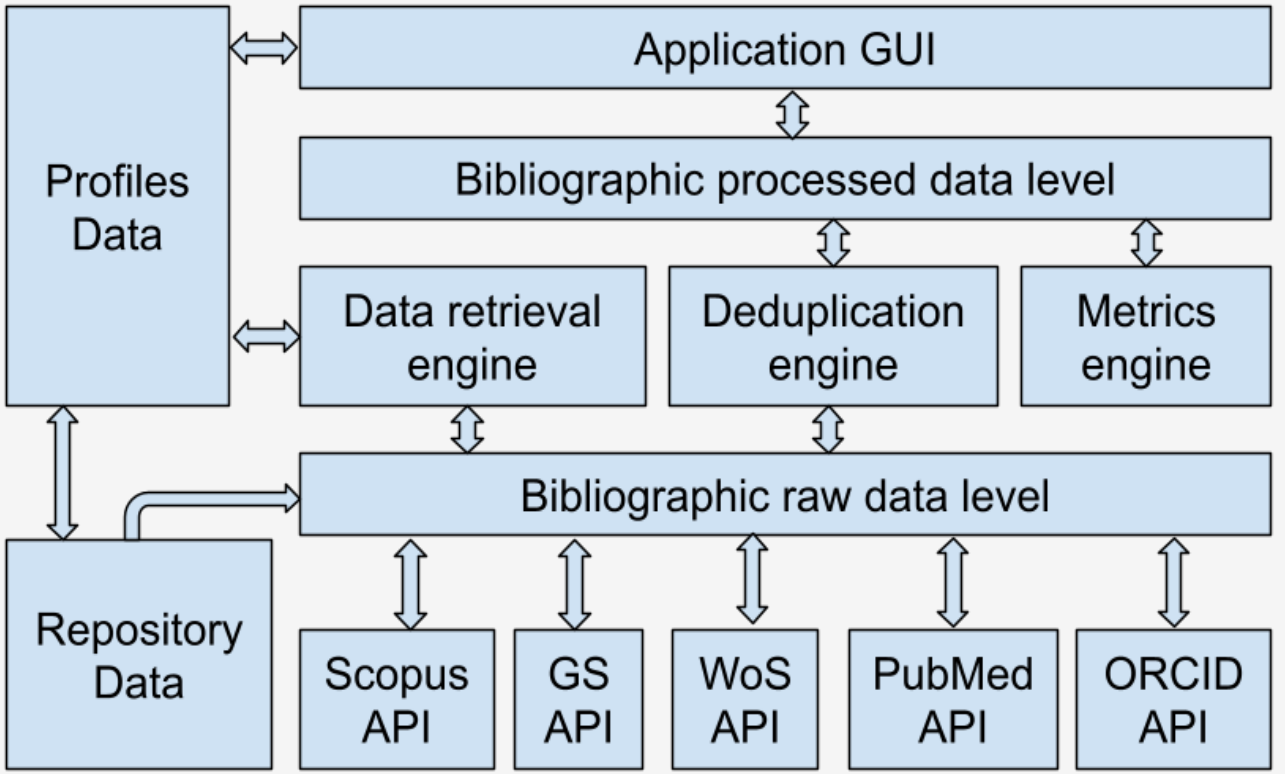
Αρχιτεκτονική της εφαρμογής - Τεχνολογίες

- **Προφίλ των φυσικών προσώπων ήδη καταχωρημένα βιβλιογραφικά δεδομένα - (Repository Data).**
- **APIs για κάθε μια από τις βάσεις αναφορών**
- **Μηχανή ανάκτησης δεδομένων (data retrieval engine)**
- **Τα δεδομένα που αντλούνται από τις βάσεις αποθηκεύονται σε κατάλληλη μορφή (bibliographic raw data level)**
- **Μηχανισμός αποδιπλοποίησης (deduplication engine)**
- **Τα δεδομένα που ενοποιούνται αποθηκεύονται σε κατάλληλη μορφή (bibliographic processed data level)**
- **Στατιστικά στοιχεία (Metrics engine)**



Αρχιτεκτονική της εφαρμογής - Τεχνολογίες

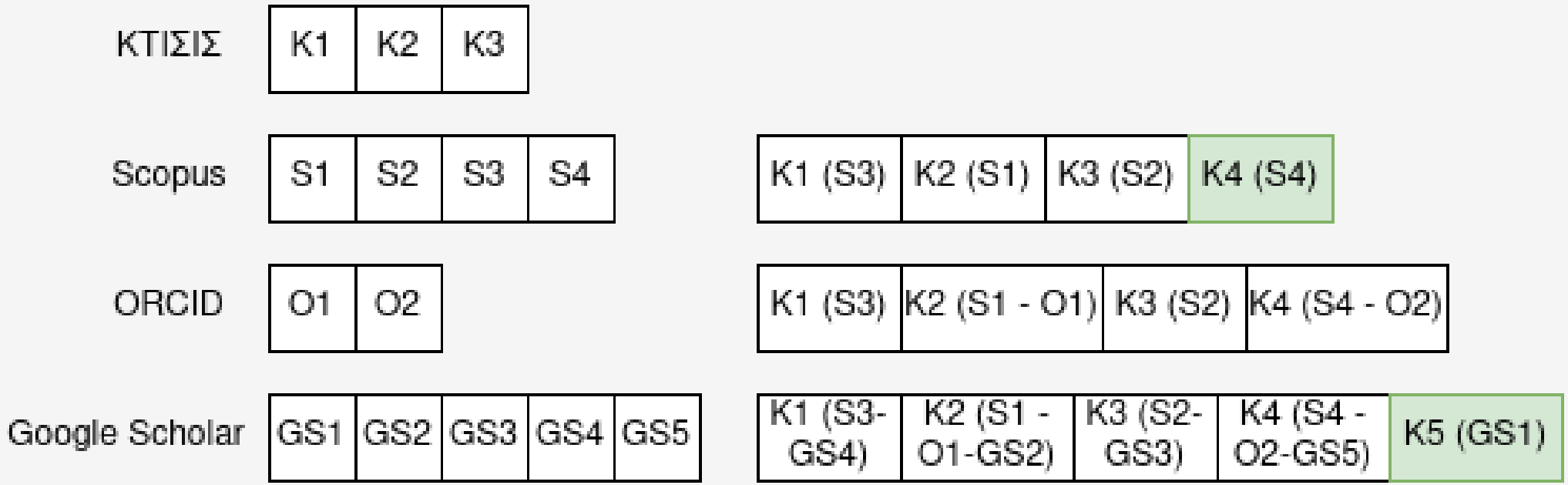
- **Για την διεπαφή εργασίας και διαχείρισης των δημοσιεύσεων αναπτύχθηκε περιβάλλον εφαρμογής με την χρήση τεχνολογιών php , postgresql , nginx**
- **Για την διασύνδεση με τα API's των πηγών από τις οποίες το σύστημα λαμβάνει τις δημοσιεύσεις αναπτύχθηκαν με την χρήση τεχνολογιών php και python .**
- **Ο μηχανισμός αποδιπλοποίησης, με κύριο γνώμονα την ανάγκη για πιθανή μελλοντική κλιμάκωση στον όγκο των δεδομένων, υλοποιήθηκε πάνω στην πλατφόρμα του Apache Spark.**
- **Τέλος το σύνολο των επιπέδων της εφαρμογής λειτουργούν εντός της τεχνολογίας docker, οπότε έτσι διασφαλίζεται η ομαλή λειτουργία και συνέχεια του συστήματος σε επεκτάσεις.**



Αποδιπλοποίηση βιβλιογραφικών δεδομένων

Ο κύριος άξονας της αποδιπλοποίησης είναι η εύρεση της ομοιότητας μεταξύ δύο δημοσιεύσεων Α και Β, έτσι ώστε, αν ο βαθμός της μεταξύ τους ομοιότητας ισούται ή είναι μεγαλύτερος από ένα προκαθορισμένο κατώφλι, να συμπεραίνουμε ότι η Β είναι διπλότυπο της Α.

Η ομοιότητα δύο δημοσιεύσεων υπολογίζεται πάνω στην κανονικοποιημένη απόσταση Levenshtein των τίτλων τους (Navarro, 2011).



Αποτελέσματα – APIs ευρετηρίων

Ευρετήριο	Συνδρομή	API	Χρήση APIs - Περιορισμοί	Μορφή δεδομένων	Αξιολόγηση δεδομένων
Scopus	Ναι	Ναι	Κλήση με βάση το Scopus ID [μοναδικό αναγνωριστικό χρήστη] και χρονικό περιορισμό	XML	**Μεγάλη κάλυψη**, **Καλή ακρίβεια** και **ορθότητα**, **πολύ καλή δόμηση** σε πεδία
Web of Science	Ναι	Ναι	Κλήση με βάση το Researcher ID [μοναδικό αναγνωριστικό χρήστη]	JSON	**Μεγάλη κάλυψη**, **Καλή ακρίβεια** και ορθότητα, πολύ καλή δόμηση σε πεδία
Google Scholar	Όχι για την πρόσβαση μέσω Web - Ναι για API	Όχι - Υπηρεσία μέσω τρίτων	Κλήση με βάση το Google Scholar Profile ID [μοναδικό αναγνωριστικό χρήστη]	JSON	**Πολύ μεγάλη κάλυψη**, Πολλά σφάλματα, μέτρια δόμηση δεδομένων
PubMed	Όχι	Ναι	Κλήση με βάση το Επώνυμο [μη ύπαρξη μοναδικών αναγνωριστικών]	XML	**Μικρή κάλυψη**, Καλή ακρίβεια και ορθότητα, μέτρια δόμηση σε πεδία
ORCID	Όχι	Ναι	Κλήση με βάση το ORCID [μοναδικό αναγνωριστικό χρήστη]	JSON	**Μεγάλη κάλυψη**, Καλή ακρίβεια και ορθότητα, πολύ καλή δόμηση σε πεδία

Αποτελέσματα – APIs ευρετηρίων

Scopus

id: "84882814680",
eid: 2-s2.0-84882814680,
title: "Integrated use of remote sensing, GIS and precipitation data for the assessment of soil erosion rate in the catchment area of \"Yialias\" in Cyprus",
name: "Atmospheric Research",
creator: "Alexakis D.",
url: https://api.elsevier.com/content/abstract/scopus_id/84882814680,
issn: "01698095",
isbn: **null**,
eissn: **null**,
volume: "131",
issue_identifier: **null**,
page_range: "108-124",
cover_date: "2013-09-01",
cover_display_date: "September 2013",
doi: 10.1016/j.atmosres.2013.02.013,
description: "The objective …..",
citation_count: "121",
med_id: **null**,
type: "Journal",
subtype: "ar",
subtype_description: "Article",
author_count: "3",

Scopus

keyword: "AHP | Cyprus | Erosion | GIS | Remote sensing | RUSLE",
source_id: "12092",
fund_acr: **null**,
fund_no: "undefined",
fund_sponsor: **null**,
open_access: "0",
open_access_flag: "0",
is_source: **null**,
last_cited_by_extraction: **null**,
created_at: "2021-10-07T13:48:40.000000Z",
updated_at: "2021-10-07T13:49:39.000000Z"

Google Scholar

id: "tDnnZQIAAAAJ:e5wmG9Sq2KIC",
title: "Integrated use of remote sensing, GIS and precipitation data for the assessment of soil erosion rate in the catchment area of “Yialias” in Cyprus",
type: "journal",
venue: "Atmospheric Research",
year: "2013",
authors: "Dimitrios D Alexakis, Diofantos G Hadjimitsis, Athos Agapiou",
publication: "Atmospheric Research 131, 108-124, 2013",
cited_by: "165",
cites_id: "14014517470617431430",

Google Scholar

link: https://www.sciencedirect.com/science/article/pii/S0169809513000744,
publication_date: "2013/9/1",
publisher: "Elsevier",
description: "The …..",
pages: "108-124",
issue: **null**,
volume: "131",
total_citations: {
table: [
{year: **2013**, citations: **3**},,
cited_by: {
link: https://scholar.google.com/scholar?oi=bibs&hl=en&cites=14014517470617431430&as_sdt=5,
total: **165**,
cites_id: "14014517470617431430",
…...},
scholar_articles: [{
link: https://scholar.google.com/scholar?oi=bibs&cluster=14014517470617431430&btnI=1&hl=en,
title: "Integrated use of remote sensing, GIS and precipitation data for the assessment of soil erosion rate in the catchment area of “Yialias” in Cyprus",
authors: "DD Alexakis, DG Hadjimitsis, A Agapiou - Atmospheric Research, 2013",
cited_by: {
link: https://scholar.google.com/scholar?oi=bibs&hl=en&cites=14014517470617431430&as_sdt=5,
total: **165**, cites_id: "14014517470617431430",
serpapi_link:https://serpapi.com/search.json?cites=14014517470617431430&engine=google_scholar&hl=en
},

Αποτελέσματα – Αποθετήριο Κτισις

Μέλη Κτίσις		Ύπαρξη Προφίλ				
Κατηγορία	**Αριθμός**	**Scopus**	**WoS**	**Google Scholar**	**PubMed**	**ORCID**
Μέλη Κτίσις	305	271	117	186	Αναζήτηση με βάση το επώνυμο	241

Ευρετήριο	**Αριθμός εγγραφών που συλλέχθηκαν**	**Άρθρα σε επιστημονικά περιοδικά**	**Ανακοινώσεις σε συνέδρια**	**Άλλο**	**Αδιευκρίνιστο**
Κτίσις	9798	5458	2757	978	605
Scopus	7250	5025	1602	620	3
Web of Science	2111	2098	0	13	0
Google Scholar	13605	6794	1983	646	4182
PubMed	565	464	0	98	3
ORCID	5946	3894	1319	733	0

Αποτελέσματα – Αποθετήριο Κτισις

Κατηγορία	Σύνολο	Scopus	WoS	Google Scholar	PubMed	ORCID
Νέα δημοσιεύματα	**5078**	**1275**	**70**	**2913**	**15**	**805**
Διπλές εγγραφές	25929	7745	2056	10295	550	5136
Προς έλεγχο	**1243**	**176**	**47**	**806**	**8**	**206**
Σύνολο εγγραφών (με διπλές σε περίπτωση συν-συγγραφέων από το Τεχν. Παν. Κύπρου)	32250	9196	2173	14014	573	6147

Καθολικοί βιβλιομετρικοί δείκτες

Πηγή	Σύνολο εγγραφών	Άρθρα σε περιοδικά	Ανακοινώσεις σε συνέδρια	Βιβλία - Κεφάλαια βιβλίων	Λοιπά / Χωρίς τύπο
Αποθετήριο Κτίσις	209 / 2 διπλότυπες	92	82	4	34
Google Scholar *	305	144	61	13	87
Πηγή	**Σύνολο εγγραφών**	**Πηγή**	**Σύνολο εγγραφών**	**Πηγή**	**Σύνολο εγγραφών**
Scopus	144	WoS	67	PubMed	3
Πηγή			**Σύνολο εγγραφών**		
ORCID			223		

Σύνολο εγγραφών προς **συγχώνευση**: **951** - **Νέες εγγραφές** σε σχέση με τις εγγραφές του αποθετηρίου Κτίσις: **81** (26 GS - 53 ORCID - 1 Scopus - 1 WoS) - **Εγγραφές που χρήζουν έλεγχο** αν τελικά είναι διπλές ή όχι: **41** - **Εγγραφές που είναι διπλές**: **661**

Από την **ενοποίηση των εγγραφών** προκύπτει μία λίστα από **288 εγγραφές** (207+81).

63 εγγραφές δεν έχουν τύπο που να ανήκει σε μια από τις κατηγορίες όπως άρθρα σε περιοδικά, ανακοινώσεις σε συνέδρια και βιβλία - κεφάλαια βιβλίων ή έχουν ελλιπή μεταδεδομένα.

Απομένουν 225 εγγραφές που ανήκουν στις κατάλληλες κατηγορίες για τον υπολογισμό βιβλιογραφικών δεικτών (118 άρθρα σε περιοδικά - 95 ανακοινώσεις σε συνέδρια - 12 βιβλία - κεφάλαια βιβλίων)

Καθολικοί βιβλιομετρικοί δείκτες

Πηγή	Σύνολο εγγραφών	Αναφορές	h-index
Scopus	144	1.887	23
WoS*	116 / 66	1,536 / 1,369	21
Google Scholar	305	2909	27
Εφαρμογή Ενοποιημένης Διαχείρισης	**225**	**1978 - 2342 - 2707 ****	**25 - 25 - 29 ****

* Η βάση αναφορών Web of Science παρέχει το σύνολο των δημοσιευμάτων / αναφορών και τον δείκτη h-index με σύμφωνα με τα ευρετηριασμένα δεδομένα του συγγραφέα (Citation network) και από άλλες πηγές και με σύμφωνα με το περιεχόμενο της Core Collection (Citation Report λειτουργία)

** Με δεδομένο ότι οι πληροφορίες για τις αναφορές ανά δημοσίευση μπορεί να προέρχονται από διαφορετικές βάσεις αναφορών για λόγους πληρότητες εμφανίζονται το άθροισμα αναφορών και δείκτης h-index με βάση την ελάχιστη τιμή αναφορών, τον μέσο όρο και την μέγιστη τιμή.

DEMO

Link Εφαρμογής

Συμπεράσματα

Η δημιουργία των εργαλείων για την άντληση βιβλιομετρικών δεδομένων αποτελεί μια πολύ σύνθετη διαδικασία που επηρεάζεται σημαντικά από παράγοντες όπως:

- Η υλοποίηση των APIs (τεχνολογία, μηνύματα, δομή δεδομένων κ.λπ.) **απαιτεί η κάθε περίπτωση να αντιμετωπιστεί διαφορετικά**.
- Τα περισσότερα από τα **APIs απαιτούν την ύπαρξη συνδρομής** (WoS, Scopus, ORCID) με το φορέα που τα διαχειρίζεται ή/και **επιβάλλουν περιορισμό στα ερωτήματα** (requests) που μπορεί να εκτελέσει κάποιος. Ειδικά για την περίπτωση του **Google Scholar δεν παρέχεται κάποιο API** από την Google και επομένως η πρόσβαση στα δεδομένα απαιτεί την **χρήση υπηρεσιών τρίτων**.
- Οποιαδήποτε **αλλαγή στο τρόπο κλήσης των APIs** αλλά και στην **οργάνωση των δεδομένων** θα απαιτήσει **επιπλέον ανάπτυξη** από την πλευρά της εφαρμογής ώστε να προσαρμοστεί στις αλλαγές.
- Για την σωστή λειτουργία είναι απαραίτητο τα **φυσικά πρόσωπα να διαθέτουν προφίλ** στα ευρετήρια αναφορών και να έχουν επιλυθεί τυχόν προβλήματα διπλότυπων εγγραφών. Επίσης είναι ακόμα σημαντικό να **υποστηρίζονται μοναδικά αναγνωριστικά (unique identifiels)** για όλες τις οντότητες που συμμετέχουν στο μοντέλο δεδομένων ανά ευρετήριο.

Συμπεράσματα

Η κάλυψη που παρέχει η κάθε βάση αναφορών είναι διαφορετική. Η **ευρύτερη κάλυψη παρέχεται από την Google Scholar** με δεδομένα όμως τα οποία **δεν έχουν υποστεί σχεδόν κανένα έλεγχο** ως προς την ποιότητά τους και την εγκυρότητά τους. Για το λόγο αυτό προτείνονται τα εξής:

- Τα **δεδομένα** που προέρχονται, κυρίως από την **Google Scholar πρέπει να ελέγχονται** πριν αυτά συμμετέχουν στην διαδικασία ενοποίησης (αποδιπλοποίησης). Ο έλεγχος πρέπει να αφορά τόσο το είδος του τεκμηρίου, τη χρονολογία έκδοσης αλλά και την ποιότητα των μεταδεδομένων.
- Η **σειρά** με βάση την οποία **θα αξιοποιούνται τα δεδομένα των ευρετηρίων αναφορών** προτείνεται να είναι η εξής: **πρώτα οι "εμπορικές" βάσεις αναφορών**, π.χ. WoS και Scopus και μετά οι βάσεις αναφορών που προκύπτουν μέσα από αυτόματες διαδικασίες δημιουργίας (π.χ. Google Scholar).

Συμπεράσματα

Για τον υπολογισμό **καθολικών βιβλιομετρικών δεικτών** για φυσικά πρόσωπα ή άλλες ακαδημαϊκές οντότητες **η προτεινόμενη εφαρμογή** μπορεί να **παρέχει μια σαφώς ακριβέστερη και πληρέστερη εικόνα** από κάθε άλλη βάση αναφορών.

Προϋπόθεση αποτελεί η εφαρμογή μιας διαδικασίας ελέγχου τόσο από πεπειραμένο προσωπικό της βιβλιοθήκης αλλά και τους ίδιους τους συγγραφείς.

Η εξαγωγή αναλυτικών βιβλιομετρικών δεικτών (π.χ. αριθμός αναφορών, h-index κ.λπ.) παρουσιάζει μια σχετική ακρίβεια αλλά **στηρίζεται στον αριθμό των αναφορών που εμφανίζει το κάθε δημοσίευμα στις επιμέρους βάσεις** και όχι **στην ύπαρξη ενός γράφου αναφορών** εντός της εφαρμογής.

Μελλοντικές επεκτάσεις

Για το μέλλον, προγραμματίζεται η βελτίωση της **παρουσίασης** των αποτελεσμάτων, η περαιτέρω **βελτίωση** του **αλγορίθμου αποδιπλοποίησης**, η **αναβάθμιση** των **διεπαφών** για την καλύτερη **άντληση** των **δεδομένων**, κυρίως μέσα από την ελαχιστοποίηση του αριθμού των κλήσεων, η προσθήκη και **νέων ευρετηρίων** (Dimensions, DataCite, Zenodo, CrossRef κ.λπ.) κ.λπ.

Μεγάλο βάρος θα δοθεί στην **δημιουργία μιας σειρά από στατιστικούς δείκτες** ανά φυσικό πρόσωπο ή ακαδημαϊκές οντότητες προσαρμοσμένες στις **απαιτήσεις των Ελληνικών Πανεπιστημίων** (συμμόρφωση με τα στοιχεία που ζητούνται στα πλαίσια της αξιολόγησης των Ιδρυμάτων).

Η εφαρμογή τέλος **θα παρέχει τα κατάλληλα APIs** τόσο για τον **εμπλουτισμό των σελίδων των Ιδρυμάτων** και των **Τμημάτων** (προφίλ καθηγητών κ.λπ.) όσο και των Ιδρυματικών Αποθετηρίων.

Τέλος παρουσίασης …

Ευχαριστούμε για την προσοχή!

Ερωτήσεις