Journal of Integrated Information Management

Vol 6, No 1 (2021)

Jan-June 2021

Citation indexes integrated management for Institutional Repositories data enrichment

Dimitris Kouis, George Veranis, Marios Zervas, Petros Artemis, Andreas Giannakopoulos, Christos Bellas, Konstantina Christopoulou

To cite this article:

Kouis, D., Veranis, G., Zervas, M., Artemis, P., Giannakopoulos, A., Bellas, C., & Christopoulou, K. (2021). Citation indexes integrated management for Institutional Repositories data enrichment. *Journal of Integrated Information Management*, *6*(1), 14–24. Retrieved from https://ejournals.epublishing.ekt.gr/index.php/jiim/article/view/37892

Journal of Integrated Information Management electronic journal

Citation indexes integrated management for Institutional Repositories data enrichment

Dimitrios Kouis[1], George Veranis[1], Marios Zervas[2], Petros Artemis[2], Andreas Giannakopoulos[1], Christos Bellas[3] Konstantina Christopoulou[1]

[1] Department of Archival, Library & Information Studies, University of West Attica, Athens, dkouis@uniwa.gr [**ORCID**: 0000-0002-5948-9766] - gveranis@gmail.com - gianandr4@gmail.com - kchristopoulou@uniwa.gr, [2]Library and Information Services, Cyprus University of Technology, marios.zervas@cut.ac.cy - petros.artemi@cut.ac.cy, [3]Aristotle University of Thessaloniki, chribell@csd.auth.gr

Article Info

Article history:
Received 5 May 2021
Received in revised form 25 May 2021
Accepted 15 June 2021

https://doi.org/10.26265/jiim.v6i1.4490

Abstract:

Purpose *– An important problem for researchers and for agencies (e.g., Quality Assurance Units) that are responsible for evaluating the research activity of academic entities (e.g., laboratories, departments, entire institutions, etc.) is to locate and retrieve the bibliographic records (e.g., scientific papers) and their citations automatically from the various citation indexes.*

Design/methodology/approach *- To calculate uniform bibliometric indicators, the deduplication of the documents collected from the different citation indexes is required. In addition, such a tool could assist the academic libraries in upgrading their Research Repositories with auto-enrichment capabilities, saving valuable labour time from their staff.*

Findings *- In this context, the initial results of implementing such a tool for data extraction from the four popular citation indexes (Scopus, Google Scholar, Web of Science and PubMed) and the ORCID service are presented. The tool aims to provide integrated management of multiple citation indexes, namely the collection of data per researcher and the application of deduplication algorithms so that a list of unique publications is obtained for each one of them. The processed data are combined with the data of the Institutional Repository and converted into a suitable format for ingestion.*

Originality/value *- The Institutional Repository of the Cyprus University of Technology has been selected as a testbed. All universities can undoubtedly utilize the obtained results.*

Index Terms — Bibliometrics, Citation Indexes, Institutional Repositories, SCOPUS, Web of Science, Google Scholar, PubMed, ORCID.

I. Introduction

A citation index is more than a simple source of bibliographic references since it provides a strict construction and a thoroughly defined data model **[1]**.

Nowadays, there are many citation indexes, such as Web of Science (Clarivate Analytics), Scopus (Elsevier), Google Scholar (Google), Microsoft Academic (Microsoft) and Dimensions (Digital Science & Research Solutions Inc.) **[2]**, as well as individual, specialized databases such as PubMed. Other services are also worth mentioning, such as ORCHID, ResearchGate etc., and unique identifier providers (PID - Persistent Identifiers) for digital objects such as CrossRef and DataCite, which develop and maintain graphs of bibliographic data.

From the very first years of the emergence of citation indexes, back in the 2000s, a series of problems came to light concerning the scientific field coverage (thematic coverage), the volume coverage (number of sources indexed), the precision of the data and the accuracy of the bibliometric indicators. Since then, hundreds of research efforts have been trying to answer the previous inquiries with interesting results.

For instance, several research papers attempt to compare and evaluate the repositories utilizing various methods. More specifically, **[3]** estimates that the balance between Google Scholar and Scopus indexes differs from 1 to 4 depending on the thematic field. **[4]** reported similar results, where Google Scholar and Microsoft Academic have the same average references values but double compared to Web of Science and Scopus. Something to keep in mind is the research statement, warning that this should not affect the authors' judgement when choosing the best citation index because many other factors influence the quality of the results, such as the calculation method.

The findings mentioned above that Google Scholar provides broader coverage in bibliographic data is confirmed by other researchers. Specifically, **[5]** realize that Google Scholar traces 95% of references from Web of Science and 92% from Scopus for all individual thematic fields. Still, at the same time, it provides almost 50% more references that are non-traceable from Web of Science and Scopus. Of course, although Google Scholar provides the most comprehensive coverage, with an estimate of 389 million records **[6],** it does not comply with the strict guidelines about what is supposed to be included in its database (i.e., it includes blogs, websites, PowerPoint files etc.) and is based mainly on crawling techniques with questionable results as far as their quality and their preciseness **[7, 8, 9]**.

Therefore, the significance of choosing the right index becomes obvious. However, the main problem is that every index returns different search results (with minor or significant deviations) for the publications and references; hence, the bibliometric values are different. In this context, it appears that there is a need for a tool that will provide unified management of bibliometric data, including the popular citation indexes, emphasizing the deduplication of identical publications. **[2]** present an elaborated overview of numerous bibliometric analysis tools. The data management process includes extracting publications from multiple sources and deduplicating them (i.e., BibExcel). Nevertheless, the most common practice is editing data from multiple sources autonomously, without any unification method provided.

According to the analysis above and to enrich the Institutional Repository Ktisis, from the Cyprus University of Technology with data from its academic staff, this paper presents the details of creating an application for the unification of bibliographic data management from sources such as Scopus, Web of Science, Google Scholar, ORCHID and PubMed.

II. METHODOLOGY

The steps presented below were followed to develop the framework to unify the bibliographic data management from multiple sources.

A. Study and production of the specifications for the interface

Choosing the databases/indexes which will participate in the bibliographic data extraction: The choice was based on criteria such as the completeness of the data, the thematic coverage and mainly on the ability to export them through Application Programming Interfaces (APIs).

Studying the available APIs and the data provided by each database/index: The target of this stage is to understand the possibilities of each API (calls, messages, fields etc.), the search options available (e.g., per individuals, per publication, etc.) and mostly the structure and the content of the results. Much attention was paid to discovering the unique identifiers for the individual entities captured in each index's data (e.g., authors, publications, institutions – affiliations, thematic fields etc.).

Creating a minimum level of common data – Data model: Each index offers different capabilities and different data fields per entity (for example, author attributes, bibliographic record attributes, etc.). Furthermore, even the record of data in similar fields (e.g., year of publication, pages etc.) follow different patterns, mainly because of the primary data provided by the publishers. The main result of this step is creating a data model, which will accommodate indexes data uniformly.

Defining deduplication algorithm of different citation indexes – bibliometric indexes: At this point, the method of the deduplication of bibliometric documents had to be specified on a scale of authors or other entities. A primary parameter in this procedure is the definition of the fields on which the algorithm will be based.

User Interface – Functions provided – Data Output: The final step concerns the user interface (UI) and the offered functions of the unified management of multiple citation indexes. The main points of this step were the way to access the application, the specific functions offered, the workflows, the statistics, the configuration parameters, etc.

Given the facts mentioned above, the main requirements and specifications were defined, signifying the basis of the software's development. Several points and choices had to be renewed/improved during the development and even more during the testing.

B. Software architecture – Technologies

The software development supporting the interface for the unified management of the citation indexes followed the logic behind an architecture like the one pictured below in the following image.

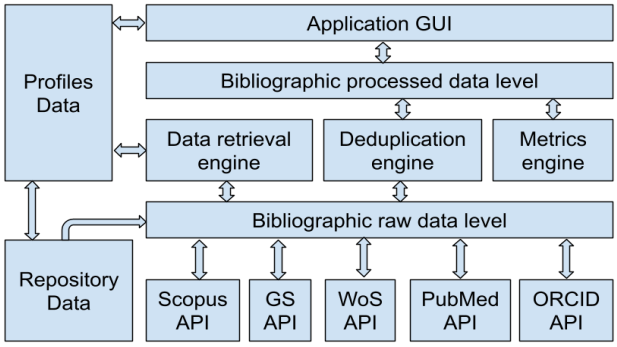

Fig. 1. The logic behind the Software's Architecture for supporting the interface

More specifically, the first step was to export the Author's profiles' data and their already registered bibliographic data provided from the Repository Data. Afterwards, the required APIs were to be developed for each citation index (Scopus, Google Scholar - GS, Web of Science - WoS and PubMed APIs) as well as for ORCHID for the extraction of bibliographic data based on the unique identifiers of the authors, such as those given in their profiles. The extracted data for every profile was saved in a specific form (bibliographic raw data level), following the same data model, as the one defined in the previous section. The data retrieval engine is responsible for the data extraction method management, providing the proper credibility mechanisms for the integrated data transfer and storing it in a specific form. The data retrieval is possible not only per user and citation index but also in a bulk mode. After completing the data extraction, the deduplication engine is responsible for the deduplication of the bibliographic records per Author per citation index and for the repository data. The deduplication engine is also responsible for identifying duplicate documents for a set of individuals (e.g., members of a department or a School). Such a capability is helpful for the creation of a unique list of publications that will be used later for the enrichment of the repository with new data. Given the great importance of the

deduplication method of bibliographic records, the next section will explain the steps of the deduplication method.

Finally, for the implementation, the following technologies and tools were utilized for each functional level:

- A software interface was designed using php, postgresql, nginx for the work interface and publication management.
- The API's interconnection of the sources from which the system retrieves the publications has been developed using php and python technologies.
- The Apache Spark platform has realized the deduplication mechanism, emphasizing the need for a potential escalation in the data mass.

The sum of the software levels works in the docker technology to ensure smooth functioning and a continuality of the system in expansions.

B. Deduplication of bibliographic data

After completing the data retrieval, the ability to start the deduplication mechanism is given. The central idea behind the deduplication process is calculating the similarity level between two publications A and B. If the level of similarity equals or is greater than a predefined threshold, we safely can assume that publications are the same. The similarity between two publications is calculated on the normalized Levenshtein distance of their titles **[11]**. Because of the squared computational complexity in finding the similarities for every publication collected, the deduplication mechanism works on two levels for the best possible workload management. On the first level, the duplicate publications per Author are calculated and the new ones appear from the data obtained by the citation indexes. In the figure given below, the 1st level function is depicted. More specifically, the algorithm receives a list of the Author's publications stored in the Institutional Repository Ktisis, gradually building the final list with the unique publications of the Author through continuous iterations, while maintaining the information for the duplicate publications that are found.

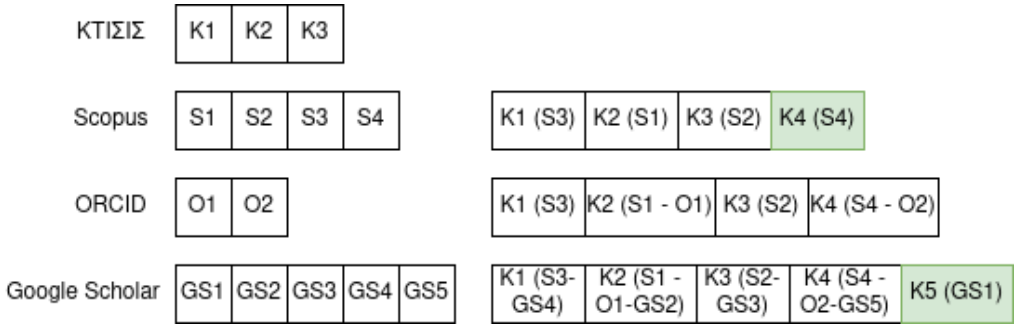

Fig. 2. Example of deduplication per Author, where the duplicates are in the parentheses and the new publications are coloured green

On the second level, the deduplication mechanism is performed on the new publications that are found for every Author, aiming the information extraction of the co-authors

III. Results

This section contains the most important results and information obtained using the application for the unified management of multiple citation indexes. The most interesting point concerns the user interface, the retrieval of bibliographic data through the APIs' and the performance of the deduplication algorithm.

A. Interface for management of multiple citation indexes

The interface for the unification of bibliographic data offers all necessary functions for achieving its purpose (see the following figure – **Basic functions**). Specifically, the user can enable the data retrieval process for all individuals per API (see the following figure – **APIs call**). Moreover, it can initiate the deduplication process for all retrieved data (see the following figure – **Run deduplication for all**). From the main dashboard, the user can access the sum of the deduplication data or per citation index (see the **Unification Statistics**).

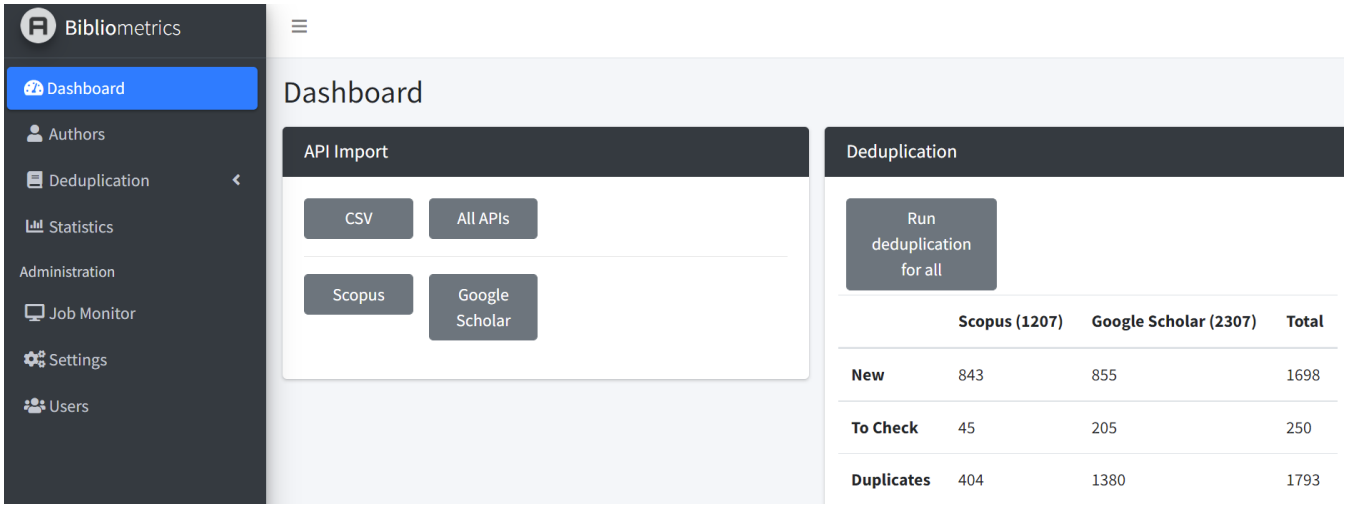

Figure 3. Main dashboard

The function offered for each person is similar. Through the Author's selection, the user can search for individuals, and after that, the user can transfer to the personal interface (**Figure 4**). Following the same strategy as in the main dashboard, the ability to call the APIs, unifying process, and statistics viewing is offered. The user can also see the records per API (as shown in Figure 4. **bibliographic data per API**).

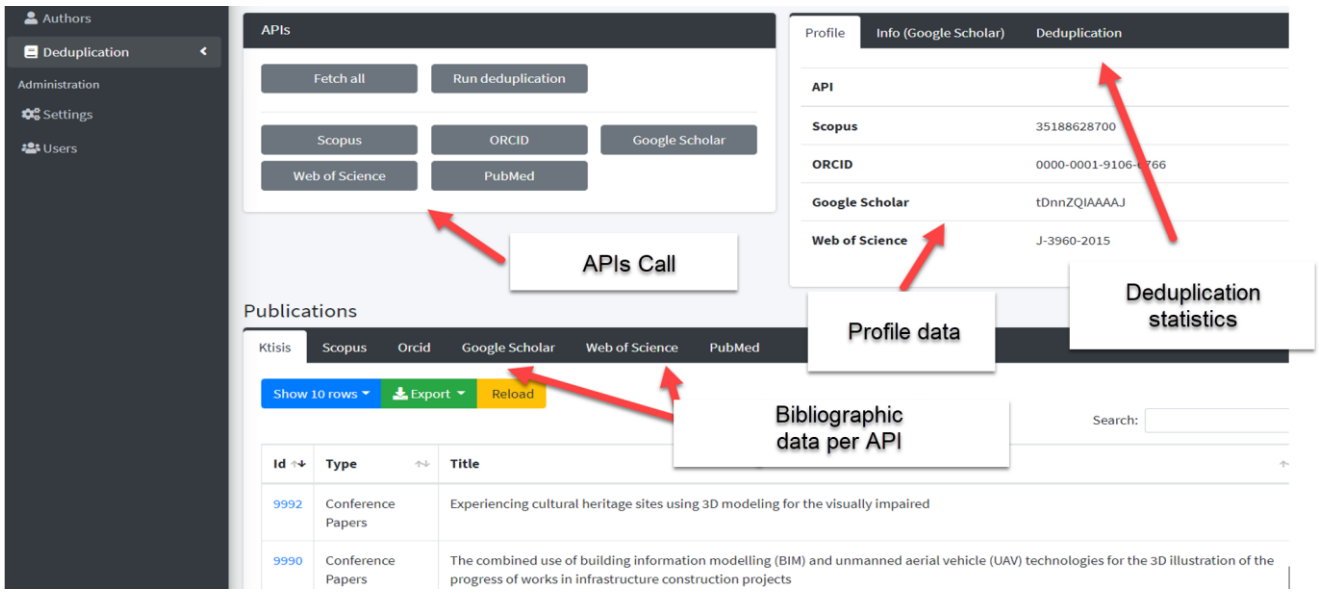

Figure 4. Interface for individuals (Authors)

An important application feature for the unified management of bibliographic data is its configuration interface. More specifically, this interface offers the ability to map the data extracted from the APIs to the specific fields of the common data model. In addition, the tolerance limits (threshold) of the data deduplication algorithm can be adjusted (Figure 5. **Deduplication**).

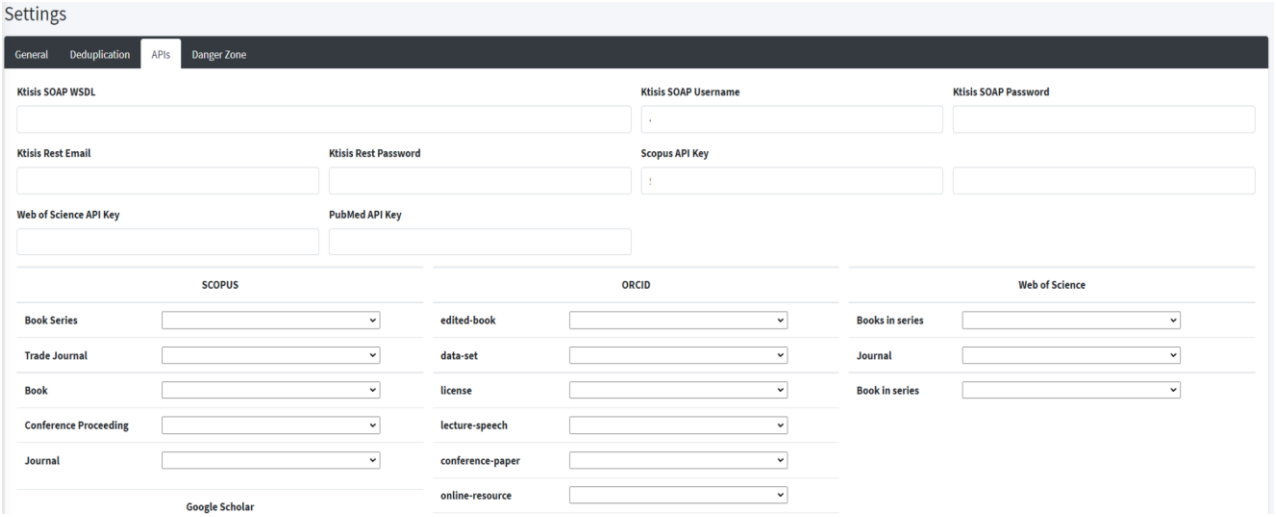

Figure 5. Deduplication

Having presented the basic functions of the unified management interface for bibliographic data, some representative results from its operation for the Ktisis Institutional Repository of the Cyprus University of Technology will be given in the following sections.

B. Developing data retrieval mechanisms for citation indexes APIs – Evaluation of the data quality

For each citation index, a separate data retrieval interface was developed. There were differences in the APIs capabilities in each case, thus at the retrieved data. The access type (free or paid) was also placing an important obstacle in each citation index's way of utilization. The thematic coverage and the precision/correctness of the given data also affected, as expected, the way of running the deduplication algorithm. The table that follows specifies the functional details of each citation index API.

Table 1. Main characteristics of the citation indexes

Citation Index	Subscription	API	Use of APIs - Restrictions	Data type	Data evaluation
Scopus	Yes	Yes	Call based on Scopus ID [unique user ID] and time limit	XML	Extended coverage, high precision and correctness, great field structure
Web of Science	Yes	Yes	Call based on Researcher ID [unique user ID]	JSON	Medium coverage, good precision and correctness, great field structure
Google Scholar	Not for Web access- Yes for APIs	No – third party service	Call based on Google Scholar Profile ID [unique user ID]	JSON	Extended coverage, Many errors, good field structure
PubMed	No	Yes	Call based on surname [lack of unique user ID]	XML	Limited coverage, good precision and correctness, good field structure
ORCID	No	Yes	Call based on ORCID [unique user ID]	JSON	Limited coverage, good precision and correctness, great field structure

According to the findings given in Table 1, a paid subscription is mandatory for three out of five citation indexes, either to their providers (WoS, Scopus, ORCID) or on third-party members, to achieve the bibliographic data retrieval process. The data included on each system were also different in the organization method (e.g., unique IDs for each individual or the lack of it) but mostly on their precision. A noteworthy finding is that for the case of Google Scholar, which offers the larger document count per individual, the data retrieved contain many errors and inconsistencies on titles, authors, document type, publication dates, etc.

The multiple versions of the same bibliographic record are given in the following figure, as found in three different citation indexes.

Scopus	Web of Science	Google Scholar

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",
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"
	id: "000323994200011",
	author_id: "J-3960-2015",
	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",
	year: "2013",
	issn: "0169-8095",
	eissn: "1873-2895",
	isbn: "1873-2895",
	doi:
	10.1016/j.atmosres.2013.02.013,
	created_at: "2021-10-09T13:38:29.000000Z",
	updated_at: "2021-10-09T13:38:53.000000Z"
		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",
		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
		},
		{
		year: 2014,
		citations: 8
		},
		{
		…….
		],
		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
		},
		versions: {
		link:
		https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=140145174706174314300,
		total: 6,
		cluster_id: "14014517470617431430",
		serpapi_link:
		https://serpapi.com/search.json?cluster=14014517470617431430&engine=google_scholar&hl=en
		},
		related_pages_link: {
		link:
		https://scholar.google.com/scholar?oi=bibs&hl=en&q=related:hm2oLmORfcIJ:scholar.google.com/
		}
		}
		],
		created_at: "2021-10-07T14:36:48.000000Z",
		updated_at: "2021-10-09T09:04:22.000000Z"
		}

Figure 6. Example of a record in different citation indexes

C. Application on the Ktisis repository

For the best possible understanding of the integrated management tool functions, its application on the Ktisis Institutional Repository of the Cyprus University of Technology will be presented. The main target of the interface is the comparison of the recorded publications on the Ktisis repository for the University staff (until a time spot) with the publications recorded in the citation indexes. If the new publications are retrieved and traced for the members

of the University was decided. The Ktisis repository publications were exported in a proper format and included in the application for each member. This way, the information contained on the Ktisis repository would remain updated and consistent. The purpose is for the information provided by the Ktisis repository to be precise and complete using the unification interface for citation index management since it will present the sum of the publications for each individual and not just a part of them. If this happens, it will be possible to extract aggregate results for both individuals and academic entities (e.g., Academic departments, laboratories, etc.).

Through the Ktisis repository, the profile data of each individual was extracted, for which the citation index management application would apply. (See Table 2).

Table 2. Ktisis members – Profile data on citation indexes

Ktisis Members		**Profile Existence**				
Category	**Count**	**Scopus**	**WoS**	**Google Scholar**	**PubMed**	**ORCID**
Members	**305**	**271**	**117**	**186**	**Search by name**	**241**

Given the information shown above, the results' accuracy is significantly affected by the existence of a profile with unique user IDs for each individual (e.g., on Pubmed the search is performed using the Author's surname). Table 3 shows the numeric values from the first level of the application (the data retrieval).

Table 3. Statistics for the data retrieval of multiple citation indexes and the Ktisis repository

Citation Index	**Records collected**	**Articles on scientific journals**	**Conference papers**	**Other**	**Undefined**
Ktisis	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

Next, the deduplication algorithm provided the following results, based on the methodology described earlier (Table 4).

Table 4. Statistics of data retrieval of the citation indexes and the Ktisis repository per individual

Category	**Sum**	**Scopus**	**WoS**	**Google Scholar**	**PubMed**	**ORCID**
New publications	5078	1275	70	2913	15	805
Double records	25929	7745	2056	10295	550	5136
To be checked	1243	176	47	806	8	206
Record sum (with duplicates in case of co-authors from the Technical University of Cyprus)	32250	9196	2173	14014	573	6147

In the final step, to create the data set to be ingested in the Ktisis repository, the duplicate records on each category were unified considering the multiplicity of the authors that happen to be members of the Cyprus University of Technology.

D. Universal bibliometric indicators

The ability to retrieve data from different citation indexes for an individual and the unification - deduplication process

offers the ability to compute the fundamental bibliometric indicators from the start with more sufficient data.

In the following table, the publication data for a member of the Ktisis repository and the record number of the documents retrieved from the rest of the citation indexes are depicted.

Table 5. Data from an individual

Source	**Record number**	**Articles on journals**	**Conference announcements**	**Books– Book chapters**	**Others/ Without a type**
Ktisis repository	209 / 2 duplicates	92	82	4	34
Google Scholar *	305	144	61	13	87
Source	**Record number**	**Source**	**Record number**	**Source**	**Record number**
Scopus	144	WoS	67	PubMed	3
Source			**Record number**		
ORCID			223		

** A closer look at the results from Google Scholar verifies the findings presented in Table 1 regarding multiple errors, even if it outperforms the other citation indexes on the record number.*

After applying the deduplication algorithm on the individual's data, the following conclusions were drawn:

- Number of records to be merged: **951**
- New records regarding the records of the Ktisis repository: **81** (26 GS - 53 ORCID - 1 Scopus - 1 WoS)
- Records that need to be checked if they are duplicates or not: **41**
- Duplicate records: **661**

Finally, there is a list of 288 records (207+81) after unifying the documents, directly correlated to the examined individual. Studying these records, some interesting facts emerge. To be more specific:

- 63 records have no type that belongs to one of the categories such as articles on journals, conference announcements and books – book chapters; or have insufficient; metadata. The important thing is that most come from Google Scholar and concern items that were "incorrectly" added to the consolidation process.
- There are 225 records with proper categories for bibliographic indicators (118 journal articles – 95 conference announcements – 12 books – book chapters).

Given these records, the following indicators – statistical values occur:

Table 6. Comparison of basic bibliometric indexes for an individual based on data retrieved by the top three citation indexes

Source	**Record number**	**Citations**	**h-index**
Scopus	144	1.887	23
WoS*	116 / 66	1,536 / 1,369	21
Google Scholar	305	2909	27
Unification Management Interface	**225**	**1978 - 2342 - 2707 ****	**25 - 25 - 29 ****

**The Web of Science database provides all publications / reports and the h-index according to the author citation data (Citation network) and from other sources and according to the content of the Core Collection (Citation Report function)*

*** As the information for the reports per publication may come from different reference databases for completeness reasons, the total number of reports and the h-index are displayed based on the minimum report value, the average and the maximum value.*

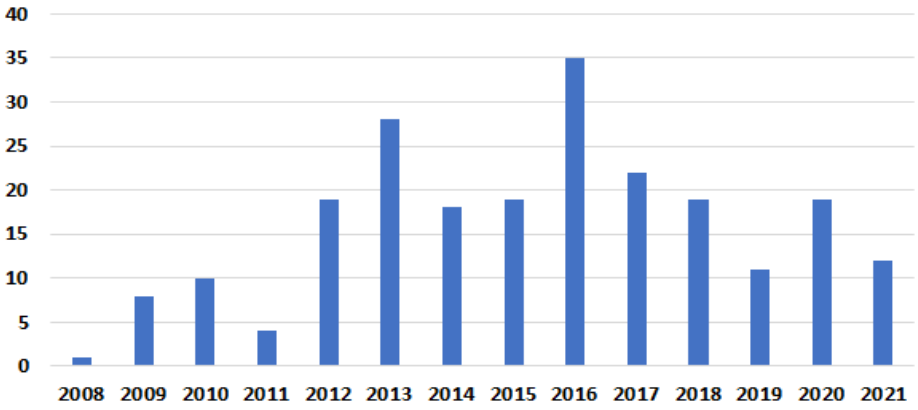

Figure 4. Number of publications per year for an individual

Based on the analysis given, the usability of the application for the unification management of citation indexes is evident since it offers a much more precise picture of the fundamental bibliometric indexes of an individual.

IV. Conclusions

The main conclusions from this attempt of creating a tool for the unified management of bibliometric data from multiple citation indexes are summarized below.

- The creation of tools for the retrieval of bibliometric data is a highly complex procedure, and it is significantly affected form factors such as:
  - The creation of the APIs (technology, messages, data structure etc.,) demands to be handled differently for each case.
  - Most of the APIs demand a paid subscription (WoS, Scopus, ORCHID) to the provider that manages or/and has restrictions on the requests that can be handled. Especially in the case of Google Scholar, there is no API provided by Google, and therefore, access to the data demands the use of third-party services.
  - Any change on the calling method of the APIs and the data organization will demand more development on the application to be adjusted to the changes.
  - For the proper function of the application, the individuals must have a profile at the citation indexes and any duplicate record issues to have been resolved. Moreover, it is still vital for all entities to support unique identifiers.
- Each citation index's coverage is different. Google Scholar achieves the broadest coverage with unchecked data for their quality and validity. Therefore, it is advised:
  - The data that come, mainly from Google Scholar, should be checked before being included in the deduplication process. The check should be on the entity type, the publication year, and the metadata quality.
  - The order in which the data of the citation indexes will be handled is: first the "commercial" citation indexes, e.g., WoS and Scopus, and then the citation indexes that derive from automatic creation procedures (e.g., Google Scholar).
- The proposed application could offer a significantly more precise and fuller picture than any given citation index to calculate global bibliometric values for individuals or other academic entities. Applying a validation procedure is a prerequisite, not only from the experienced library staff but also from the authors themselves. The extraction of analytical bibliometric values (e.g., number of citations, h-index, etc.,) shows some preciseness but it is based on the number of citations of each publication as given in the citation indexes and not on a reference graph.
- The data deduplication / unification algorithm presents great results and it can be easily adjusted. It's interface (Apache Spark) allows a future escalation.

The improvement of the presentation of the results, the further improvement of the deduplication algorithm, the alteration of the APIs for a more efficient data retrieval primarily by minimizing the repetition calls, the addition of new citation indexes (Dimensions, DataCite, Zenodo, CrossRef etc.,) and so on, is scheduled for the future. Emphasis will be given on creating a series of statistical indicators per individual or academic entities; based on the needs of Greek Universities (compliance with data required for their evaluation). In conclusion, the application will provide the proper APIs for the enrichment of the websites of the Institutions, the Departments (professors' profiles, etc.) and the Institutional Repositories.

V. References

[1] McVeigh, Marie E. 2017. "Citation Indexes and the Web of Science." Encyclopedia of Library and Information Sciences. 4. Edition. Edited by John D. McDonald and Michael Levine-Clark. Boca Raton London New York: CRC Press, vol. 2: 940-50.

[2] Moral-Muñoz, José A.; Herrera-Viedma, Enrique; Santisteban-Espejo, Antonio; Cobo, Manuel J. (2020). “Software tools for conducting bibliometric analysis in science: An up-to-date review”. El profesional de la información, v. 29, n. 1, e290103. https://doi.org/10.3145/epi.2020.ene.03

[3] Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. Journal of Informetrics, 10(2), 533–551. https://doi.org/10.1016/j.joi.2016.04.017

[4] Michael Levine-Clark & Esther L. Gil (2021) A new comparative citation analysis: Google Scholar, Microsoft Academic, Scopus,

and Web of Science, Journal of Business & Finance Librarianship, 26:1-2, 145-165 163, https://doi.org/10.1080/08963568.2021.1916724

[5] Martín-Martín, Alberto, Enrique Orduna-Malea, Mike Thelwall and Emilio Delgado López-Cózar. 2018. "Google Scholar, Web of Science, and Scopus: A Systematic Comparison of Citations in 252 Subject Categories". Journal of Informetrics 12, no. 4: 1160–77.

[6] Gusenbauer, Michael. 2019. "Google Scholar to Overshadow Them All? Comparing the Sizes of 12 Academic Search Engines and Bibliographic Databases". Scientometrics 118, no. 1: 177–214. https://doi.org/10.1007/s11192-018-2958-5

[7] Harzing, Anne-Wil, and Satu Alakangas. 2016. "Google Scholar, Scopus and the Web of Science: A Longitudinal and Cross-Disciplinary Comparison." Scientometrics 106, no. 2: 787-804. https://doi.org/10.1007/s11192-015-1798-9

[8] Harzing, Anne-Wil. 2019. "Two New Kids on the Block: How do Crossref and Dimensions Compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science?" Scientometrics 120, no. 1: 341-9. https://doi.org/10.1007/s11192-019-03114-y

[9] Sugimoto, Cassidy R. and Vincent Larivière. 2018. Measuring Research: What Everyone Needs to Know. Oxford: Oxford University Press.

[10] Navarro Gonzalo. 2001. A guided tour to approximate string matching. ACM Comput. Surv. 33, 1 (March 2001), 31–88. DOI: https://doi.org/10.1145/375360.375365

VI. AUTHORS

Dimitrios Kouis received his Diploma in Computer Engineering and Informatics from the University of Patras and his PhD from National Technical University of Athens (NTUA) in 1994 and 2004 respectively. His scientific interests include Library Networks, Digital Publishing, Scholarly Communication topics, Software development, Content Management, IT middleware platforms, meta-data modelling etc. He has been involved in several European and national projects and has published more than 30 articles in journals and conferences. Currently, he is an assistant professor at the Department of Archival, Library and Information Studies, University of West Attica.

George Veranis received his Diploma in Applied Informatics and his MSc in Information Systems from the University of Macedonia in Thessaloniki in 2005 and 2008 respectively.

He has worked as Senior Software Developer for Hellenic Academic Libraries Link Services since 2008. He has been involved in numerous projects developing, deploying and enhancing several applications, systems and websites to meet and satisfy objectives. He is also very experienced with interoperability between different applications because he has developed a middleware based on that and has participated in other open-source projects. Moreover, has developed open source smILLe which is used by National Network of Interlibrary Loan in Greece. His research interests include knowledge management, could computing, information retrieval, software engineering, semantic web technologies, data mining, bibliometrics and research evaluation methods, library management systems, web services and middleware systems. He has 12 published articles in international scientific journals and conference proceedings.

Marios Zervas received his diploma in Genie Informatique from the Centre National des Arts et Métiers of Paris, France, and his MBA from the Open University of Cyprus in 1994 and 2015. His scientific interests include Open Access, Information literacy, School libraries, and Cultural Heritage. He is the Coordinator of the Erasmus + Project Educability "Building the Capacity of Librarians and Educators in Information Literacy," a member of IFLA School Libraries Committee, and the Vice President of Cyprus Libraries Consortium. He is the Library Director at the Cyprus University of Technology

Petros Artemi holds a BSc in Library Science & Information Systems, a MA in Digital Video Production and a M.Sc. in Social Information Systems. He works as a system administrator at the Library of the Cyprus University of Technology since 2007. For more than 12 years, he is deeply involved in different digitalization projects and have experience with open-source software, metadata schemas and data models. His research interests focus on institutional repositories and linked data.

Andreas Giannakopoulos received his Bachelor degree in Electrical and Electronic Engineering from the University of West Attica in 2019. He has various interests in the Information Technology field such as artificial intelligence, software development, web development, robotics, internet of things, data science, bibliometrics etc. He is also a cloud practitioner certified with the "DevOps Engineer Expert" in the Microsoft Azure platform. Currently his role is "Automation and DevOps Engineer" at Performance Technologies and he also involved with a bibliometrics project for the University of West Attica.

Konstantina Christopoulou has earned her Bachelor's and Master's Degree in Computer Science from the University of the Peloponnese and is a PhD student in the Department of Archival, Library and Information Studies, University of West Attica. Her scientific interests include Bibliometrics, Data Visualization, Data Science, Education, Cultural Informatics etc. She has been a member of Knowledge and Uncertainty Research Laboratory of University of Peloponnese, hosting multiple educational programs and seminars, while being in charge of organizational duties such as press releases and members'

data handling. She is currently a member of the Information Management Research lab of University of West Attica, contributing as a teaching assistant and a researcher.