Towards a Universal and Interoperable Scientific Data Model

João Oliveira
University of Minho, Guimarães, Portugal

Diogo Gomes
University of Minho, Guimarães, Portugal

Francisca Santana
University of Minho, Guimarães, Portugal

Jorge Oliveira e Sá
Algoritmi Centre, University of Minho, Guimarães, Portugal

Filipe Portela ✉
Algoritmi Centre, University of Minho, Guimarães, Portugal
IOTECH - Innovation on Technology, Trofa, Portugal

Abstract

The growing number of researchers in Portugal has intensified the appearance of several scientific platforms that allow the indexation of publications and the management of scientific profiles. The diversity and high number of platforms brings problems at the level of crossover and integrity of the information, i.e., the researchers' profiles are rarely updated, and their data are not properly grouped and cross-referenced. Hence, the need arises for a more global platform that enables the synchronization of information, free from constraints imposed by existing data.The study and work carried out aims to solve this problem by creating a robust and interoperable platform based on an innovative library merge algorithm. Thus, this platform includes information regarding publications, researchers and scientific indicators, by crossing and grouping data from several platforms.

2012 ACM Subject Classification Software and its engineering → Development frameworks and environments

Keywords and phrases RDProfile, Researchers, Scientific Platforms, Scientific Data

Digital Object Identifier 10.4230/OASIcs.SLATE.2023.14

Funding This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

1 Introduction

Nowadays, the number of existing scientific data, libraries, metrics, and indicators (KPIs) is enormous and makes it difficult to manage them by researchers and scientific institutions. The number of researchers has also increased, which makes it even more difficult to access correct and updated information on multiple platforms. Therefore it is vital to collect and combine scientific data from several platforms to answer the institution's requests and improve the quality of project applications by enhancing the researcher's data. To this end, it is necessary to create an algorithm for merging libraries, as part of the RDProfile project – a project that aims to create an inclusive, reliable, and scalable solution that can easily include diverse information from the profiles of researchers.

In the scope of this study, the following research question can be identified throughout the article: How can the data from profiles, quartiles, and indexes from various platforms be conciliated?

The main objective is to develop an artifact that includes a prototype of a web solution capable of grouping and cross-referencing data from researchers from various sources including Scopus, Web Of Science, and ORCID. To accomplish this goal, it is necessary to create secondary objectives such as developing a responsive interface with pervasive features, developing an API that allows communication with some of the platforms, integrating several APIs, and optimizing RDProfile platform with three new features (profiles, quartiles, and index).

This paper is divided into six sections: Introduction, Background, Material and Method, RDProfile Algorithm, Proof of Concept, and Conclusion. In the first chapter, the theme of the paper is contextualized, as well as its objectives and research question are defined. The most important concepts are described in Background. The Scrum and DSR methodologies are covered in the next chapter – Material and Method – as well as the used tools and metrics, and the three scientific platforms involved in the work, namely, ORCID, Scopus and Web Of Science. In the fourth chapter, RDProfile Algorithm, the architecture and structure of the RDProfile platform is described, with focus on the Middleware (API), where the developed library merging algorithm fits. In Proof of Concept concrete cases are presented, using real data from a Portuguese research institution and finally, in the last chapter, Conclusion, final considerations are made and future work is described.

2 Background

In this section the most important concepts are covered, in order to better understand the background of this paper. Thus, the project to which this work belongs is described, as well as the general concept of scientific platforms and the related work.

2.1 RDProfile

RDProfile is a complete and interoperable platform that allows grouping and cross-referencing researchers' data from several sources, such as Scopus, ORCID and Web Of Science. The goal is to provide users with correct and updated data of researchers and institutions, acting as an aggregator of dispersed knowledge, publicly available and easily accessible. Besides being an inclusive, reliable and scalable solution, RDProfile also intends, through gamification, to help identify the best scientific profiles and improve indicators that allow increasing the number of R&D projects funded based on the improvement of the quality of researchers' metrics [10].

The RDProfile Project started in 2020 and has three phases [10]. The first phase – Research and support the project idea – resulted in an understanding of the state of the art and the definition of the requirements for the RDProfile, providing a clear vision of the added value it adds. The second phase – Explore a possible solution to solve the problem – will result in the creation of a library merging algorithm that will solve the problems that currently exist, such as incorrect identification of references. Afterward, a web solution will be implemented, which will have innovative and useful functions for scientific researchers, such bibliometric indicators about researchers and institutions, metrics about projects and reviews, filters that allow easy cross-referencing, monitoring of citations lists with indexations, among others.

In the third phase, the artefact will be optimized and made universally available. Here the focus will be on ensuring scalability and modularity so that the solution can be accessed by any institution by Interface or API.

2.2 Scientific Platforms

The expansion of Internet has also expanded the number of digital platforms with similar objectives, such as Web Of Science and Scopus, which have been the two most widely used databases for bibliometric analyses [13]. We can divide these platforms into two groups. The indexing platforms allow the indexing and cataloging of articles in a scientific database, and the aggregating platforms, which aggregate various data in order to offer the user more information about publications, researchers and institutions. RDProfile, as already mentioned, is an aggregator since its goal includes making it easier and more efficient to search for researchers and institutions, providing the majority of data in one place.

According to previous study conducted as part of the RDProfile project [2], the most relevant scientific platforms are Scopus, Clarivate Web Of Science, ORCID, Authenticus, Google Scholar, Crossref, Scholarpedia, Semantic Scholar and Dimensions. However, the work will focus on the first four due to their similarity to the RDProfile platform.

Overall, these platforms are commonly used and recognized, but they have limitations or errors, e.g., few features/data, non-matching data, missing or wrong metrics, wrong data grouping, missing citations, and so on witch prevents institutions from getting accurate data [10]. Of the four platforms, Authenticus and ORCID are aggregators, like RDProfile. Authenticus combines the data of other platforms like Scopus and Web Of Science, but it is nationally focused, not scalable, and lacks a responsive, user-centred interface. ORCID has become increasingly popular and has recorded a steady growth in ORCID registrations [1]. This platform allows looking at how individual researchers interact and produce their work. Unlike the others platforms, the researcher becomes the center of the analysis, shifting attention from merely counting publications and citations to a much richer perspective related to the scientific workforce and its internal dynamics [5]. Web Of Science and Scopus stand out as the two most traditional sources of scientometric data [5]. Scopus is often considered as one of the largest curated databases [12]. Although, it does not cover all publications, nor does it expose all metrics of either publications or authors. The same happens with Web Of Science that remains relevant for the scientific area [12]. The combination of those three relevant platforms can provide a good solution to the problem identified in this paper. The more platforms a solution covers, the more complete it becomes, however, in this stage, only these three will be considered, to avoid increasing unnecessarily complexity.

2.3 Related Work

Within the range of platforms previously identified as similar to RDProfile, a study "A Benchmarking Study of the Scientific Platforms"[3] was conducted to Authenticus platform for being the most similar platform.

Authenticus is a database of scientific publications authored by researchers from Portuguese institutions. This database aggregates publications from different scientific platforms, such as Scopus, Google Scholar, Clarivate Web Of Science, Crossref, DBLP and ORCID. Its operation is based mainly on an algorithm capable of identifying the name of an author in a publication and associating it correctly with a particular researcher in the database. Subsequently, and if the association is successful, the publication of that same researcher will be added to the database.

This study concluded that Authenticus contains some limitations, especially when it comes to crossing the data and indicators. Some of the identified limitations are the absence of the h-index indicator and quartiles per publication which are not totally correct, it does not cross the platform data with the publications coming from Google Scholar, it contains limiting filters, it only contains a visualization form and it does not contemplate indicators related to projects and reviews.

3 Material and method

Given the complexity of the work, the Design Research Science and Scrum methodologies were used, as they provided a set of principles, practices and techniques that assisted in the development of the project. Next, Design Research Science is described as a scientific methodology and SCRUM as a methodology to assist in efficient project management. Tools, relevant metrics and the three scientific platforms involved in the work, namely, ORCID, Scopus, and Web Of Science are also described in this chapter.

3.1 Design Science Research

The Design Science Research methodology has emerged in the field of information technology supported by results, which offers a set of procedures and practices for the development of research projects. Design Science Research divides the process into six steps, these being as follows:

- **Problem Identification and Motivation:** in this step, the problem and the research question are defined, and the value of the solution is justified. In the case of the present work, the problem is related to the lack of data transversality among the existing platforms, so the research question is "In what way is it possible to reconcile the data of profiles, quartiles and indexes from various platforms?".
- **Define the Objectives for a Solution:** in this step, the objectives of the solution are defined.The main goal of the solution is to group and cross-reference data from researchers from various sources (Scopus, Google Scholar, ORCID) on a global scale.
- **Design and Development:** in this work, this step defines the artifact design of the RDProfile prototype, referred to in chapter 4.
- **Demonstration:** this step presupposes tests on the functioning of the developed product, demonstrated in chapter 5.
- **Evaluation:** in this step, the developed prototype is evaluated according to the defined goals.
- **Communication:** this step includes communicating the problem and the importance of the artifact, as well as its usefulness and effectiveness to researchers and other relevant target audiences. This article is an example of the communication of the artifacts developed.

3.2 SCRUM

The Scrum is an agile methodology that uses an iterative and incremental approach to optimize predictability and to control risk, relying on three pillars of implementation, namely transparency, inspection, and adaptation [11]. The fundamental unit of scrum is a scrum team. Scrum teams, composed by teams and organizations generating value through adaptive solutions to complex problems, are cross-functional, which means that members have all the skills needed to create value every sprint [11]. Thus, the Scrum methodology was used to assist in the efficient management and structuring of the project, speeding up the development of the practical part of the work developed.

3.3 Tools

Table 1 describes the tools and python libraries used to create the algorithm.

Table 1 Tools Description.

Tools	Description
Python	Versatile language for software development, web apps, and data analysis. Provides libraries like NumPy, SciPy, and Pandas for numerical computing, scientific tasks, and data manipulation
os	Python library used to specify the working directory
glob	Python library used to return all csv files (files with data regarding quartiles)
pandas	Python library used to manipulate data, in this case to combine all csv files into one
csv	Python library to load the generated csv (combined csv)
json	Python library used to transform the combined csv into a json file, used later to load the data into MongoDB
MongoDB	Scalable and flexible non-relational database program for storing large amounts of data. Utilized for storing relevant data in the context of scientific publications and researchers

3.4 Metrics

In order to expose the most relevant information about each researcher, metrics and indicators were selected to enrich the profile of each researcher.

- **Number of Total Citations:** Number of times the researcher's scientific publications were cited by other publications;
- **H-index:** This metric is defined as the number of papers with citation number>h, as a useful index to characterize the scientific output of a researcher [8];
- **CiteScore:** It results from dividing the number of citations received in the last four years by the number of publications that have been published in that same time interval;
- **SCImago Journal Ranking (SJR):** Is a size-independent metric aimed at measuring the current "average prestige per paper" of journals for use in research evaluation processes[7]. There are four quartiles: Q1, Q2, Q3, Q4. In quartile 1 (Q1) are the most important and prestigious scientific journals, while in quartile 4 (Q4) are the least prestigious scientific journals;
- **Source normalized impact per paper (SNIP):** Aiming to allow direct comparison of sources in different subject areas, this metric results from the ratio of the scientific journal's citation count per article and the citation potential in its subject field;
- **Journal Impact Factor (JIF):** Corresponds to the average number of times articles published in the last two years have been cited in the Journal Citations Report (JCR).

3.5 ORCID

ORCID is a platform that, through a digital identifier (ORCID iD), allows the distinction among researchers. Each researcher has his ORCID iD and, in addition, also has his own register (ORCID Record) which contains information about himself and his work. Each researcher manages his own ORCID record data and may place different information about himself. This information includes his name, biography, emails, websites and social links, keywords, countries, and activities[9]. All activities from the investigator profile include:

- **Employment:** Employment lists of organizations where the researcher has been professionally affiliated;
- **Education and Qualifications:** Where the researcher studied and the educational educational or professional qualifications that were awarded to him/her;
- **Invited positions and distinctions:** Positions the researcher has held and awards or awards he or she has received in recognition of his or her achievements;
- **Membership and service:** Memberships in society or association and donations of time or other resources in the service of an organization;
- **Funding:** Grants, awards, and other funding that the researcher has received to support his or her project;
- **Works:** The researcher's work and images of his or her research results, such as publications, conference presentations, data sets among others.

3.6 Libraries

For the creation of the library merging algorithm, the Scopus and Web Of Science APIs, among others, were used. For this reason, it is relevant to understand these two indexing libraries.

Scopus is a comprehensive database specializing in abstracts and citations with enriched data and scholarly literature linked across a wide range of disciplines. This scientific platform is owned by Elsevier and covers various academic disciplines such as science, technology, medicine, social sciences and humanities. Scopus has a large number of indexed publishers and is one of the largest databases of abstracts and citations of peer-reviewed literature[6].

The Web Of Science, owned by Clarivate, is a global publisher-independent citation database. This scientific platform provides researchers with information and technology for the global scientific research community. It provides data and analytics, as well as customized tools and professional services to researchers and the entire research community such as universities, research institutions, national and local governments, and private and public research funding organizations around the world. It also supports more than 95 percent of the world's leading research institutions, multiple governments, and national research agencies with about 20 million researchers in more than 7,000 research organizations[4].

4 RDProfile Algorithm

This section is about the architecture and structure used for the development of the RDProfile platform.

4.1 Architecture

As can be seen in Figure 1, the API is communicating regularly with the APIs of ORCID, APIs, collecting and uploading data about researchers and publications to the database. The communication between the various APIs is done through functions that are timed to run weekly, allowing the information to be kept up to date. In addition, two distinct processes are carried out that aim to collect data relative to the scientific indicators (quartiles, h-index, sjr, jif, among others). The first process consists in exporting data through the SCImago platform – the rankings and quartiles from all journals since 1999 until now (2022) – , which is then transformed and uploaded to MongoDB. For the second process, a Web Scraping mechanism is used that allows collecting data from Clarivate, which is also later uploaded to the database.

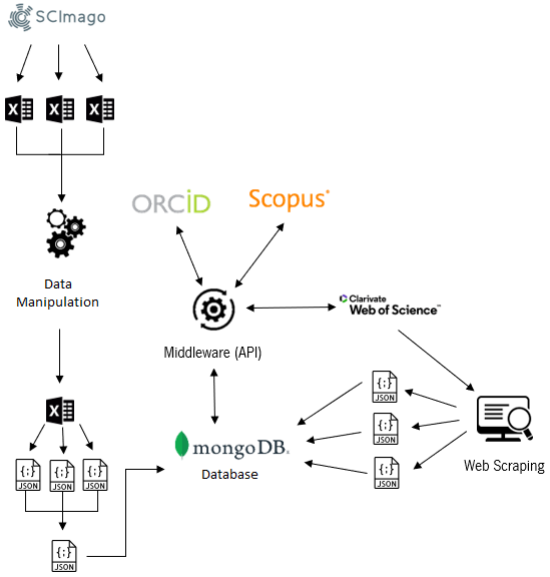

Figure 1 Architecture of the algorithm.

Figure 2 shows how the algorithm works at a high level. First it is checked if the publications are in the database. If they are, it is checked to see if they are unique, otherwise the process ends. If it is unique it is inserted in the "newPublications" array, if not it is alerted that it already exists and verifies the publications again. Next, a series of validations are performed. If the publication is from Web Of Science and Scopus it is inserted into the Web Of Science and Scopus Model, if it is from Scopus it is inserted into the Scopus Model, if it is from Web Of Science it is inserted into the Web Of Science Model. For all these validations the quartiles of the respective publications are added. If all three decisions are false, new publications are checked again and the process is repeated. All the models used are described in subsection 4.2.

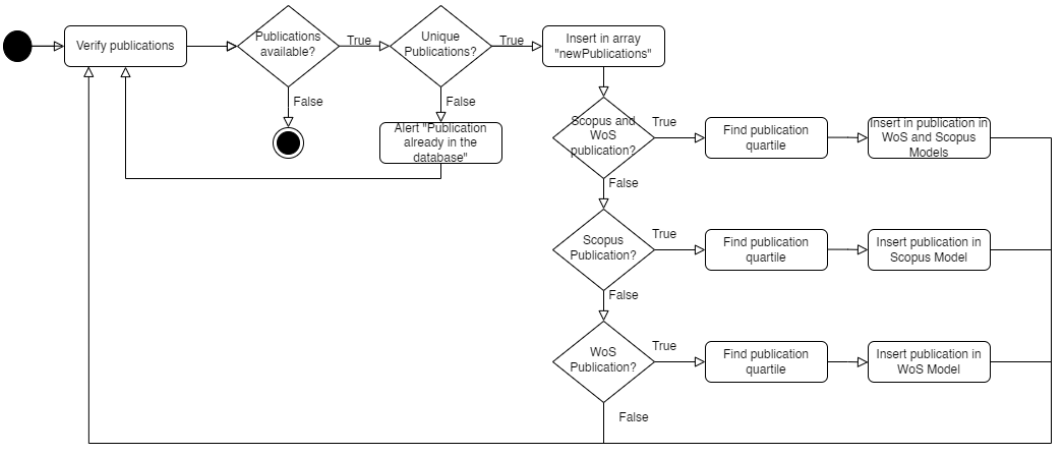

Figure 2 High level diagram of the algorithm.

4.2 Modeling, Collecting and Loading Data

The modeling phase consists of creating a data model to define how the data is related to each other and stored in the database. In total six different models were created, namely Researcher Model, Scopus Model, Web Of Science Model, Quartiles Model, Publications Model and Metrics Model. For data collection, some sources were used, namely, API of the Algoritmi Center, ORCID API, Scopus API and the Web Of Science API. After the data collection, the data from each model is loaded into MongoDB. Each created model is described below.

4.2.1 Researcher Model

The Researchers Model contains the data regarding scientific researchers. Table 2 contains a brief individual description of each field and the types of data present in this model.

Table 2 Researcher Model.

Field name	Description
id	Unique Identifier of the researcher
academic_degree	Academic degree of the researcher
degois	Unique Identifier CiênciaVitae
gscholar	Google Scholar Unique Identifier
institution	Educational institution of the researcher
name	Name of researcher
orcid	ORCID Unique Identifier
researcher_id	Web Of Science Unique Identifier
scopus	Scopus Unique Identifier
articles	Researcher's articles
name_slug	Researcher's name slug

4.2.2 Scopus Model

The Scopus Model includes data related to Scopus publications, collected using the Scopus API, which returns the scientific publication data according to the eid of the publication. Thus, the Table 3 shows all the field names and its description.

4.2.3 Web Of Science Model

The Web Of Science Model includes data related to Web Of Science publications, collected using the Web Of Science API, which returns the scientific publication data according to the uid of the publication. Thus, the Table 4 shows all the field names and its description.

4.2.4 Quartiles Model

The data from the Quartiles Model were collected through two distinct processes, referring to indicators from Scopus and Web Of Science. For the collection of Scopus indicators the following activities are performed: export the indicator data, transform the data, create the function to upload the data, and upload the data to the database. Regarding the collection of Scopus indicators, it is necessary to process the data from the Journal Citation Reports using a Web Scraping engine. Then the data are loaded into the database. Table 5 contains a brief individual description of each field present in this model.

Table 3 Scopus Model.

Field name	Description
dc_identifier	Object identifier of the document
affiliation	Publication's affiliation
article_number	Article number
author	Author of the publication
citedby_count	Number of citations of the publication
dc_creator	Lead author of the publication
eid	Unique Identifier for a publication present in Scopus
fa	Total access
link	Open access status
openaccess	Open Access status
openaccessFlag	Open Access status (boolean)
pii	Publication item identifier
prism_aggregationType	Type of aggregation (i.e. journal)
prism_coverDate	Publication date (YYYY-MMDD)
prism_coverDisplayDate	Date of publication (original text)
prism_doi	Object identifier of the document
prism_eIssn	Electronic serial number (international standard) of a publication
prism_issn	Serial number (international standard) of a publication
prism_pageRange	Interval of publication pages
prism_publicationName	Publication name
prism_url	Publication URL
prism_volume	Volume number of the publication
pubmed_id	Medline Unique Identifier
source_id	Identifier of publication's source
subtype	Subtype of publication (i.e. cp)
subtypeDescription	Description of the subtype of publication (i.e. conference paper)
title	Article title

Table 4 Web Of Science Model.

Field name	Description
uid	Unique identifier of a publication in the Web Of Science
citations	Number of citations
identifiers	Publication identifiers (doi, issn, eissn, isbn, eisbn)
keywords	Article keywords
links	Publication links
names	Publication authors
source	Origin of the publication
title	Publication title
types	Publication type (article, editorial material, meeting)

Table 5 Quartiles Model.

Field name	Description
rank	Journal rank
sourceId	Journal Unique Identifier
title	Journal Title
type	Type of journal
issn	Serial number (international standard) of the journal
eIssn	Electronic serial number (international standard) of a publication
sjr	Scimago Journal Rank
sjr_best_quartile	Best quartile of the journal
h_index	Journal h-index
jif	Journal Impact Factor
jci	Journal Citation Indicator
total_docs	Total number of documents
total_docs_3	Total number of documents in the last 3 years
total_refs	Total references
total_cites_3	Total Citations in the last 3 years years
total_citations	Total citations
citable_docs_3	Number of documents citable in the last 3 years
cites_by_doc_2	Number of citations per document in the last 2 years
ref_by_doc	Total references per document
country	Journal Country
region	Journal Region
publisher	Journal Publisher
coverage	Journal Coverage
year	Journal year
categories	Journal categories

4.2.5 Publications Model

This model includes data from Scopus and Web Of Science publications, as well as indicators related to them. The mechanism for data collection regarding the Publications Model results from the interaction between the previously loaded Scopus Model, Web Of Science Model and Quartiles Model. Table 6 contains a brief individual description of each field present in this model.

4.2.6 Metrics Model

The Metric Model includes data related to the indicators of scientific researchers, and just like the Researchers' Model resulted in a model that aggregates data from other models. The mechanism to collect data regarding the Metrics Model results from the interaction between the previously loaded Researchers Model and Publications Model. It also requests the Scopus API to obtain the researcher's h-index. Table 7 contains a brief individual description of each field present in this model.

Table 6 Publication Model.

Field name	Description
data	Identifiers (eid, uid, issn, eissn, doi, pii) and year of publication publication
dc_identifier	Object identifier of the document
title	Publication title
citedby_count	Number of citations of the publication
subtypeDescription	Description of the subtype of publication (i.e. conference paper)
link	Link to the publication
types	Publication type (article, editorial material, meeting)
scopus_id	Scopus Unique Identifier
wos_id	Web Of Science's Unique Identifier
author	Publication authors
quartiles	Publication quartiles
citations	Number of citations present in the publication

Table 7 Metrics Model.

Field name	Description
id	Unique identifier of the researcher
calgId	Algoritmi Center's Unique Identifier
academic_degree	Researcher's academic degree
orcidId	ORCID's Unique Identifier
scopusId	Scopus's Unique Identifier
gscholar	Google Scholar's Unique Identifier
researcher_id	Web Of Science's Unique Identifier
name	Researcher name
research_lab	Research Laboratory name
research_groups	Research groups name
articles	Investigator articles
citation_count	Researcher's number of citations
editorial_count	Researcher's number of editorials
h_index	H-index of researcher (Scopus)
publication_count	Researcher's number of publications
name_slug	Researcher's name slug
quartiles	Researcher's number of quartiles (total and by years)
final_score	Final score of researcher
ranking	Researcher Rank

5 Proof of Concept

The RDProfile emerges intending to create a robust solution that includes information about publications, researchers and scientific indicators, by crossing and grouping data from various platforms. Prior to the creation of the algorithm described in this paper, Algoritmi Center, a Portuguese scientific institution, developed Algoritmi Center API within the RDProfile project. Thus, the tests performed were done using real data from this institution.

To better understand the algorithm, the actual flow of the general process is shown below:

1. Collect investigator data through the ORCID API and Algoritmi center API;
2. Upload the data to the Researchers Model;
3. Collect data for each publication via Scopus API and Web Of Science API;
4. Upload publications with uid and eid simultaneously to the Publication Model;
5. Upload publications with eid to the Scopus Model and publications with uid to the Web Of Science Model;
6. Collect and load data from Scopus and Web Of Science indicators into the Quartiles Model;
7. Upload data to the Metrics Model, cross referencing data from the Researchers Model and the Publications Model.

Once the flow is complete, it is possible to see the loaded models in MongoDB. Next, real data examples of each model's result is represented in tables. Since the Metrics Model corresponds to the general model, which aggregates the data from the others models, the final result represents the aggregation of relevant data for each researcher. Because of that, it is shown, in image format, real data related with one of the authors of this paper.

In the Table 8 it is shown an example of a Researcher Model, from the researcher referred above, with its attributes.

Table 8 Researchers Model Example.

Field name	Variable
id	"730"
academic_degree	"PhD"
dgois	"F311-7C27-F8DA"
gscholar	"HoeD9UgAAAAJ"
institution	"Escola de Engenharia da Universidade do Minho"
name	"Carlos Filipe da Silva Portela"
orcid	"0000-0003-2181-6837"
researcher_id	"G-5324-2012"
scopus	[0: "57194071672"]
articles	Array
name_slug	"carlos-filipe-da-silva-portela"

In the Table 9 it is shown the data of the Scopus publication with eid 2-s2.0-85129325287, present in Scopus Model.

Table 9 Scopus Model Example.

Fields	Variable
dc_identifier	"SCOPUS_ID:85129325287"
affiliation	Array
article_number	null
author	Array
citedby_count	"0"
dc_creator	"Azevedo J."
eid	"2-s2.0-85129325287"
fa	true
link	Array
openaccess	"0"
openaccessFlag	flase
pii	null
prism_aggregationType	"Book Series"
prism_coverDate	"2022-01-01"
prism_coverDisplayDate	"2022"
prism_doi	"10.1007/978-981-16-7618-5_27"
prism_eIssn	"23673389"
prism_issn	"23673370"
prism_pageRange	"307-318"
prism_publicationName	"Lecture Notes in Networks and Systems"
prism_url	https://api.elsevier.com/content/abstract/scopus_id/85129325287
prism_volume	"350"
pubmed_id	null
source_id	"21100901469"
subtype	"cp"
subtypeDescription	"Conference Paper"
title	"Convolutional Neural Network – A Practical Case Study"

In the Table 10 it is shown the data of the Wos publication with uid 000676684500001.

Table 10 Web Of Science Model Example.

Field name	Variable
uid	"WOS:000676684500001"
citations	{ 0: { bd:"WOS", count:1, id:6393e85271c0384234a1218a} }
identifiers	{ doi:"10.3390/fi13070178"; issn:null; eissn:"1999-5903"; isbn:null; eisbn:null}
keywords	{authorKeywordsArray}
links	Array
names	{Authors Array, Book Editor Array}
source	{sourceTitle: "FUTURE INTERNET", publishYear: 2021, ... }
title	"Data Science and Knowledge Discovery"
types	{0:"Editorial material"}

In the Table 11 it is shown an example of Quartiles Model with its attributes.
In the Table 12 it is shown an example of Publiations Model with its attributes.

Table 11 Quartiles Model Example.

Field name	Variable
rank	“1”
sourceId	“16810”
title	“Annual Review of Biochemistry”
type	[0: journal]
issn	[0: 15454509, 1: 00664154]
eIssn	Array
sjr	“50,518”
sjr_best_quartile	“Q1”
h_index	“293”
jif	null
jci	null
total_docs	“30”
total_docs_3	“80”
total_refs	“5913”
total_cites_3	“3484”
total_citations	null
citable_docs_3	“80”
cites_by_doc_2	“35,78”
ref_by_doc	“197,10”
country	“United States”
region	“Northern America”
publisher	[0:“Annual Reviews Inc.”]
coverage	“1946-1948, 1950-1960, 1962-2020”
year	1999
categories	[{_id: 63596f8c8a7fc80c14936350, area:“Biochemistry”, quartil:“Q1”}]

Table 12 Publications Model Example.

Field name	Variable
data	{eid: “2-s2.0-85091404413”, uid: null, issn: null, eissn: null, doi: null, pii: null, year: 2020}
dc_identifier	“SCOPUS_ID:85091404413”
title	“A SWOT analysis of big data in healthcare”
citedby_count	“0”
subtypeDescription	“Conference Paper”
link	https://www.scopus.com/record/display.uri?eid=2-s2.0-85091404413&origin=resultslist&sort=plf-f
types	Array
scopus_id	“635018a59e663c3c83f5931a”
author	Array
quartiles	{jif: null, jcr: Array, sci: Array}
citations	Array

In the figure 3 it is shown an example of Metrics Model loaded in MongoBD. This figure shows all the relevant real data from the researcher Carlos Filipe Portela. Here we can see, among others, the researcher’s unique identifiers of Algoritmi Center, ORCID, Scopus,

Google Scholar and Web Of Science. In addition to the number of publications and citations, a list of these publications, containing all relevant data, is also included. As for the quartiles parameter, it contains, for each year, the number of publications per quartile – from quartile 1 to 4, and without quartile. Additionally, this list contains a last global element, that contains the total sum of publications by quartiles, as can be seen in the figure.

Figure 3 Metric Model example.

The algorithm developed for RDProfile fills some of the identified gaps of the existing algorithms, namely in Authenticus. This algorithm is more accurate, scalable and modular than the existing and mentioned in section 2. It was created so that any researcher can easily access their updated data. In other words, the RDProfile algorithm, in addition to the total and per-year citations and quartiles indicators, also provides the h-index indicator and indicators about projects and reviews. It also cross-references and groups all researchers' data and indicators on a global scale. Based on this, the platform will have unique features that provide a better experience for researchers and other users.

6 Conclusion

Based on the work, the feasibility of creating the library merging algorithm is verified, as well as the added value it offers. To answer the research question, a literature review was conducted and a prototype of a web solution capable of grouping and cross-referencing data from ORCID, Scopus and Web Of Science was developed. The most in-depth study of ORCID

and Authentics allowed us to understand how RDProfile can be an added value. Both the limitations of the missing indicators and the functionalities of each platform were overcome. In other words, the RDProfile platform answers the research question by integrating:

- 4 APIs (ORCID, Scopus, Web Of Science and Algoritmi Center);
- 3 endpoints of the quartiles;
- 2 endpoints for the list of researchers with and without filter applied;
- 3 endpoints for the list of publications with and without filter applied;
- 1 endpoint for the researcher profile with the associated metrics;
- a user interface, for the most recent publications;
- 13 functionalities at the API and user interface level.

The developed algorithm proves that it is possible to integrate the main sources of scientific information in a single platform with highly interoperable and cross-reference data capabilities. Therefore, it becomes possible to create a scalable solution that any researcher can use to do their own analyses, representing a contribution to science.

Further improvements in functionalities and in the user interface are needed to make RDProfile a more complete and useful platform. It is important to optimize the solution so that it includes more platforms, such as Google Scholar, Crossref, Dimensions, among others. In addition, errors related to number of citations should be corrected, since the existing systems are limited to check articles that are correctly identified. Thus, this improvement is important to ensure maximum data integrity and reliability.

References

1 Miriam Baglioni, Paolo Manghi, Andrea Mannocci, and Alessia Bardi. We can make a better use of orcid: Five observed misapplications. *Data Science Journal*, December 2021. doi:10.5334/dsj-2021-038.

2 Joanna G. Carvalho. O cv científico dos investigadores. Master's thesis, U. Minho, 2020.

3 Joanna G Carvalho, João Oliveira, Luciana Machado, and Filipe Portela. A benchmarking study of the scientific platforms. submmited to publication, 2023.

4 Clarivate Analytics. Web of science. https://clarivate.com/products/webofscience/.

5 Rodrigo Costas, Carmen Corona, and Nicolas Robinson-Garcia. Could orcid play a key role in meta-research? discussing new analytical possibilities to study the dynamics of science and scientists, May 2022. doi:10.31235/osf.io/sjck6.

6 Elsevier. Scopus. URL: https://www.elsevier.com/solutions/scopus.

7 Borja González-Pereira, Guerrero-Bote Vicente, and Felix Moya-Anegon. A new approach to the metric of journals' scientific prestige: the sjr indicator. *J. Informetrics*, 4:379–391, July 2010. doi:10.1016/j.joi.2010.03.002.

8 J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005. doi:10.1073/pnas.0507655102.

9 ORCID. What is orcid? https://info.orcid.org/what-is-orcid/.

10 F. Portela. Rdprofile - ficha tecnica, 2022.

11 Ken Schwaber and Jeff Sutherland. Scrum: A framework for managing agile projects. Online, 2013. URL: https://www.scrum.org/resources/what-is-scrum.

12 Vivek Singh, Prashasti Singh, Mousumi Karmakar, Jacqueline Leta, and Philipp Mayr. The journal coverage of web of science, scopus and dimensions: A comparative analysis. *Scientometrics*, 126, March 2021.

13 Vivek Kumar Singh, Prashasti Singh, Mousumi Karmakar, Jacqueline Leta, and Philipp Mayr. The journal coverage of web of science, scopus and dimensions: A comparative analysis. *Scientometrics*, 126(6):5113–5142, 2021.