publications

MDPI

Article

Knowledge Maps as Support Tool for Managing Scientific Competences: A Case Study at a Portuguese Research Institute

João Génio, Alina Trifan and António J. R. Neves *

IEETA/DETI, University of Aveiro, 3810-193 Aveiro, Portugal; joaogenio@ua.pt (J.G.)
* Correspondence: an@ua.pt

Abstract: In a research organization, finding someone who is an expert in a field and that can take up a given role, defining areas of excellence, or employing a new member all require understanding the competences that are available in-house. This work explores the idea of using knowledge or competence maps as support tools for managing scientific competences. We implemented a use case at the Institute of Electronics and Informatics Engineering of Aveiro, a research institute at the University of Aveiro, but the methodology we proposed can be adapted to virtually any research organization. Knowledge maps are visual representations of information that can be designed with variable granularities with respect to the knowledge assets of an organization. From a research management perspective, knowledge maps support the discovery of research competences and provide an instant overview of a topic by showing the main areas at a glance. This solution explored in this work employed data mining approaches for gathering information from public databases and presenting it using knowledge maps. Other visualization tools, such as bar graphs, tables, filters and search functionalities, were created and integrated into a web platform. When put together, these components could turn the platform into a key component for the administration of a research organization.

Keywords: knowledge maps; concept maps; network graphs; natural language processing; scientific competence management; data mining; Institute of Electronics and Informatics Engineering of Aveiro

Citation: Génio J.; Trifan, A.; Neves, A.J.R. Knowledge Maps as Support Tools for Managing Scientific Competences: A Case Study at a Portuguese Research Institute. *Publications* **2023**, *11*, 19. https://doi.org/10.3390/publications11010019

Academic Editor: Blanca Rodríguez-Bravo

Received: 28 December 2022
Revised: 2 March 2023
Accepted: 16 March 2023
Published: 21 March 2023

1. Introduction

Managing the knowledge and competences of the collaborators at an organization is a very relevant task [1–3]. From a research manager's perspective, the need to find an expert to act as the leading investigator of a research plan or as reviewer of a given research proposal could be supported by an appropriate knowledge management tool. Ultimately, the intellectual capital of an organization could be improved if this tool could act as a knowledge management enabler [4].

Among the many statistics related to science, called scientometrics, bibliometrics can be defined as a quantitative analysis of academic publishing, making it one of the few subfields concerned with measuring scientific output [5,6]. At the article level, one can calculate how many times the article has been cited by another work, which is dependent on the size of the indexing database used. Similarly, the importance of a journal in a given field can also be calculated. Finally, at the author level, one can find more complex metrics such as the popular h-index [7], that finds h publications with at least h citations.

Knowledge maps are diagrams that represent ideas with nodes and links. They are often used as media for learning activities, lectures and study materials [8]. They can be distinguished by the use of labeled nodes denoting concepts and links denoting relationships among them [9]. Efficient knowledge maps are expected to bring value to those interested in knowledge management of intellectual assets. Not only do they support the discovery of organizational knowledge, they also facilitate the interaction with

outsider stakeholders, as working knowledge maps are often seen as a clear sign of an organization's competence.

The main objective of this work was to employ data mining approaches to develop a platform that supports knowledge maps designed to assist the management of research units. This was achieved through the creation of dynamic collaboration networks [10,11] in the shape of knowledge maps.

This document is structured as follows. Section 3 explains the technologies used in this work and how they relate to each other. Section 4 goes through the process of collecting, processing and saving data from public sources. Section 5 focuses on the visual tools that were developed and how they aid in the knowledge extraction process. Section 6 presents a summary of this work and its conclusions.

2. Related Work

There are several open-source platforms available for creating and using knowledge maps. The Florida Institute for Human & Machine Cognition (IHMC)'s CmapServer is an open-source platform for creating, sharing and managing concept maps and knowledge models [12]. It includes a range of features for managing access to maps, sharing maps online and collaborating with others. It allows users to create hierarchical concept maps, cross-linked concept maps and knowledge models that can be shared online or offline. The Visual Understanding Environment (VUE) application is an open-source platform for creating and sharing knowledge maps, concept maps, and other types of visualizations [13]. It is designed to be used by educators, researchers, and students, and includes a range of features for organizing and visualizing information. MindMup is a free, open-source platform for creating and sharing mind maps, concept maps and other types of visualizations [14]. It includes a range of features for collaborating with others, exporting and importing data and customizing the appearance of visualizations. These platforms are all free to use and provide a range of features for creating and sharing different types of map visualizations, including knowledge maps. They can be used by individuals, teams or organizations to manage scientific competences and support knowledge management in a range of contexts. However, these are generic platforms that were developed under the assumption that the underlying information will be manually introduced.

With respect to similar platforms that manage scientific knowledge, we identified several solutions, which were not directly applicable to our use case, as detailed next.

Authenticus [15] is a project that was developed at the University of Porto that aimed to build a national repository for metadata of publications that were authored by researchers of Portuguese institutions. Similar to this proposed work, the system automatically imports publications from multiple indexing databases, such as *ORCID*, and conducts a redundancy or duplicate checking process [16]. Its development started in 2010, spanning beyond 2015 through a master's dissertation [17], but it has not been further developed in the last few years.

Open Knowledge Maps [18] presents to the user a topical overview based on the 100 most relevant documents matching a given query. It uses text similarity to group documents together and create the knowledge maps. It intends to give the users a head start on their scholarly search. Its main goal is to identify relevant areas at a glance and documents related to them. Its main sources are the Public Library of Science[1] and PubMed[2]. It employs natural Language processing techniques to build the knowledge maps [19]. This platform's data sources are its main limitation, which makes it unsuitable for the necessities of our use case.

Elsevier's Pure [20] is another research information management system. It comprises several features, amongst which is the extraction of data from numerous sources, and it supports workflow improvements for both researchers and institutions. *Pure*'s main disadvantage is that it is not a free tool, therefore, ruling it out for our use case.

Many Universities around the world start building portals to showcase their research activities. However, as far as we were able to gather, such platforms are not open source, therefore, they are not directly accessible for use by other institutions.

3. Architecture

Gathering information about the researchers in an institution can be cumbersome, which is why the main objective is being able to quickly find people that specialize in a certain field. One component of the proposed system is composed by external data sources. The first data source is *Elsevier*[3] for its ability to index data from other publishers and that it has the full content in their open access documents. *Ciência Vitae*[4] was also chosen because it is possible to manually add content to it, such as projects and much richer personal information. It was crucial that the extracted data could be saved locally, as public application programming interfaces (APIs) impose temporal limits for accessing information. Those data could then be presented to the user in the shape of knowledge maps, charts and tables.

The core of the system is composed of a web application implemented in *Django*[5], a *Python*[6] web framework for developing secure and scalable production applications. Finally, this web platform had to provide visualization tools that aided the research manager in its tasks. These tools were built with the help of the *chart.js*[7] and *vis.js*[8] frameworks.

Figure 1 represents the entire system at a high level. The flow of information starts with the extraction of data that are stored in a local database, processed by the web framework's backend and displayed to the user with the aid of visualization tools. Ultimately, the user has the ability to perceive the original data in a new, refined way that allows for extracting scientific knowledge from a set of researchers.

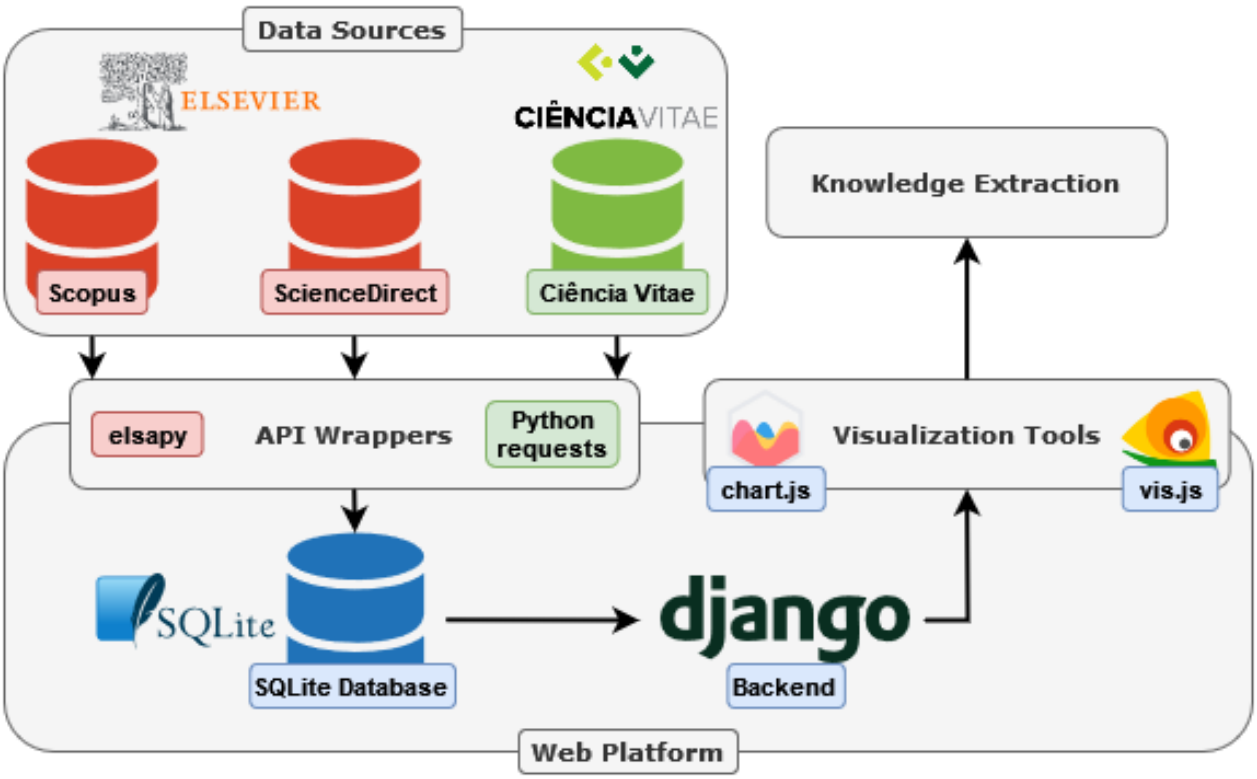

Figure 1. Main components of the web-based system.

4. Data Collection

To interact with Elsevier's APIs, we used the *Python* library *elsapy*[9]. Both the *Scopus*[10] [21] and *ScienceDirect*[11] [22] interactive APIs provided the necessary documentation, as well as

visual tools, that helped in the understanding of the data model and its structure. There is no publicly available *Python* package or library for accessing the *Ciência Vitae* API [23], so the solution was implemented on the many high-quality and robust *Python* frameworks for executing requests on the web. We ended up choosing the *requests*[12] library, as it is one of the most used libraries and it could easily satisfy our needs.

The collection process started with the identification of all researchers of a given organization and the collection of their research IDs, more specifically, their *Scopus* and *Ciência Vitae* IDs. Following that, their profiles were fetched from the APIs along with their publications. During the later stages of development, this synchronization process was transformed into a periodic task, while still allowing the user to do it manually. Every publication had to go through a duplicate checking process, otherwise the local database would end up containing redundant data, which would negatively impact the knowledge maps and, consequently, the knowledge extraction performed by the user.

4.1. Duplicate Handling

Data collection was implemented to extract extensive amounts of information from both APIs. However, this led to duplicate publications residing in the database. One of the main challenges when building a bibliometric database is the handling of redundant information at the author [24] and publication levels. Only the latter applied to this project since it was assumed that the organization's administration correctly collected and inserted the researchers' IDs into this platform.

This problem was approached from two perspectives. The first method was to check the publications' IDs, which could be a *Scopus* ID, a *Ciência Vitae* ID or a digital object identifier. A pair of publications was deemed a duplicate if they shared at least one ID. The second approach took place when the previous one did not detect a duplicate. It focused on processing and comparing both publications' titles and abstracts, through the use of natural language processing (NLP) techniques, and evaluated their similarity. Finally, the duplicate publication was merged into the existing one by joining their fields, such as keywords and scientific areas.

Title and Abstract Analysis

When we wanted to compare the titles or abstracts of two publications, we had to take into consideration many factors, including special or upper case characters, among others. This process fell into the domain of NLP, which focuses on giving computers the ability to understand text the same way human beings can. This is what we wanted to achieve in this process. In many of the pipeline stages, we used one of the most popular *Python* NLP libraries, *NLTK*[13], for its ease of use and detailed documentation.

Figure 2 describes the necessary processing that we needed to do in order to compare two strings (titles or abstracts), that were different in their character structure but equal in meaning. It also includes a publication's title as an example.

Some existing processing pipelines transform the string into lower case at a later stage, but, in this specific application, it did not change the result and it is the first step that one usually thinks of when designing a pipeline. The second step separates words that are connected by special characters, which is performed using static character replacement, and replaces all dashes with a space. "Tokenization" focuses on separating a string into smaller pieces called "tokens". This sets up the next steps for removing non alphanumeric tokens and, finally, stop words are removed, such as "to" and "the". The example in Figure 2 shows how we transformed the title "Ontology-based health information search: Application to the neurological disease domain" to "ontology based health information search application neurological disease domain". Note that this can also be applied to abstracts, and it can result in a much more efficient comparison later on, due to the removal of stop words and other elements.

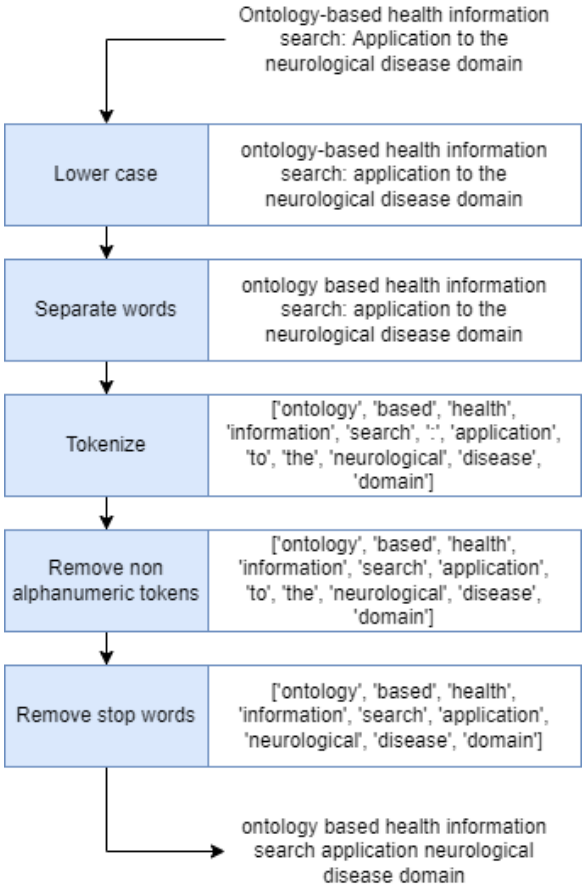

Figure 2. Text processing pipeline execution before comparing publications' titles and abstracts.

4.2. Collection Results

This work's case study revolved around a set of 50 researchers from the research institute that hosts the authors of this article. Their number of publications ranged from 12 to 312, and there were many collaborations between them. Note that the data collection process iterated through one author's profile at a time. This resulted in publications with an X amount of authors (that were present in the database) to be fetched X times from an API. A total of $X - 1$ publications were detected as duplicates because of their IDs.

In summary, there were a total of 7803 publications that were fetched from the authors' profiles, 4369 of which were added as new publications and 3434 were merged into existing ones. When looking at the merged publications we could conclude that 3103 were merged because of their IDs and 331 were merged because of the title and abstract analysis process. Additionally, Table 1 shows the number of publications with a given amount of authors.

Table 1. Publications with a given amount of authors in the database, after the collection phase.

Number of Authors	Number of Publications
1	3457
2	660
3	196
4	38
5	12
6	5
7	0
8	1

A total of 11,964 different keywords were gathered from these publications, as well as 262 scientific areas that were associated either with the publications or directly with the authors' profiles. Finally, 144 projects were fetched from the authors' *Ciência Vitae* profiles.

5. Knowledge Extraction

The knowledge extraction phase focuses on building visual tools to help extract knowledge from the gathered data. Knowledge maps were implemented as collaboration networks that could be relative to a given author or to the entire research organization. The first type of map represents which colleagues had publications in common with that author, as well as presenting the number of collaborations between all of these authors. The second type was created for the entire organization. This resulted in a disconnected graph, as a specific set of authors had not yet collaborated with the rest of their colleagues. Additionally, a data viewer was implemented to reveal the publications that each node or edge refers to. Ultimately, the user has the ability to select an author (node) or relation (edge), by interacting with the knowledge map, and view a visualization of a list of publications that belong to that highlighted element.

Filters were also introduced, as they help the map focus only on relevant information. A date range can be defined, as well as the type of publication (book, article, etc.). Finally, a search engine was introduced to modify the map to a set of keywords. When the user queries the system for an expression, it returns a set of keywords related to that query. That set can be further modified by the user, making the map more accurate to the user's needs. This search functionality was implemented by reusing the text analysis process presented in Section 4.1, for checking which keywords are similar to the user query.

Other support tools were implemented, such as customizable bar graphs that show specific types of publications in a date range, most common keywords or scientific areas for a specific author and global statistics. The total number of publications and keywords from the organization, or from an individual author, is also shown.

The platform will be available in the near future at the webpage of the Institute of Electronics and Informatics Engineering of Aveiro[14]. The project is available as an open source project at GitHub[15].

5.1. Author Map

The first attempt at relating authors and publications came in the shape of a collaboration map of a specific author. The main objective was to find the authors that someone collaborated with and the number of collaborations. However, this information alone can be represented in a table and still be easy to interpret, so, in order to make the map useful, the relations were expanded to the collaborators. This meant that the collaborations between the author's colleagues were also represented, giving the user a new layer of information to view.

Figure 3 represents a map that was generated with the collected data. This map focuses on a specific author and their colleagues, making it a collaboration map. It also indicates that, in a given set of conditions, the strongest relation occurred between the author and one of their colleagues, with five publications in common (collaborations).

5.2. Global Map

The global map is an extension of the author map to an institutional level. It is not built around any specific author, but, instead, it evaluates all authors present in the database. This gives the viewer a global perspective about the scientific knowledge that the institute possesses. Figure 4 represents a map with the same parameters as Figure 3, but relative to the entire research institute. It is possible to observe that the map in Figure 3 is included in this one.

Both the authors and relations from the author map that was presented before are shown, along with new information on other authors. The previous map now appears as it has been expanded to the colleagues' collaborations with other authors that did not collaborate with the main author from Figure 3 (in these specific conditions). The other

observation is that it is now possible to observe authors that have publications that match the criteria but did not collaborate with anyone else, making them isolated from the rest of the map. There is also a pair of authors that collaborated only with each other, on the left of the image.

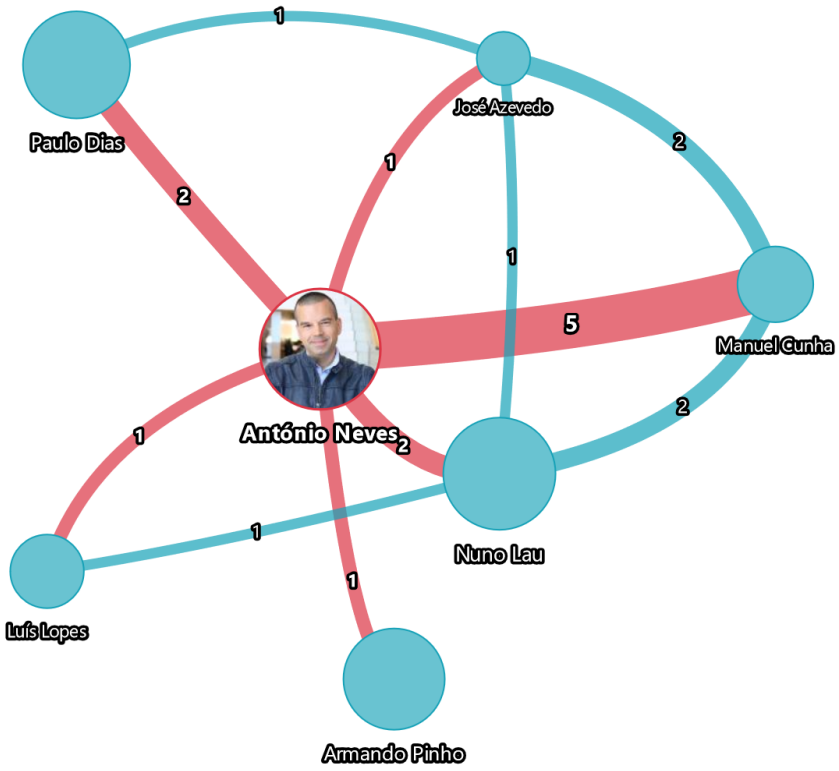

Figure 3. Example of an author's collaboration map.

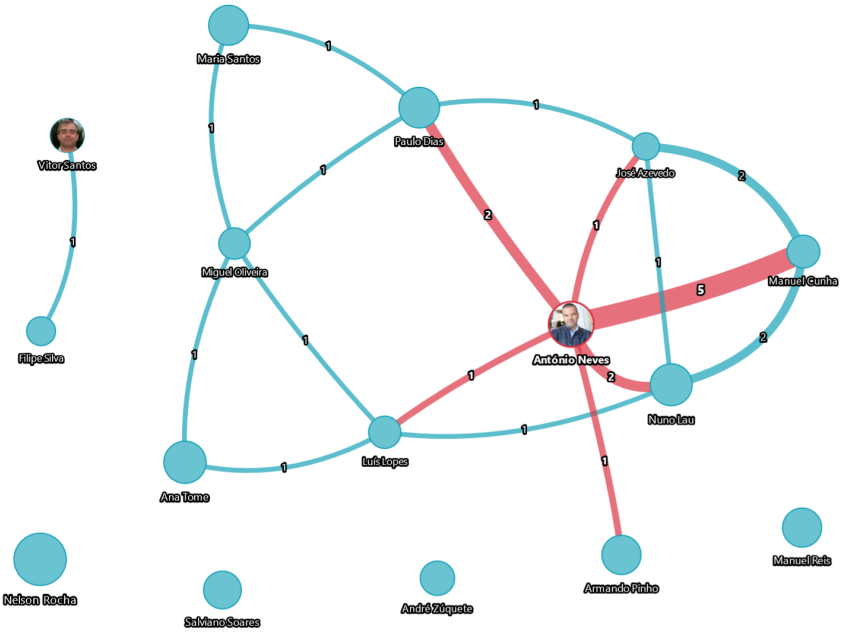

Figure 4. Example of an institute's collaboration map.

5.3. Search and Filter Functionalities

In addition to the knowledge maps, a tool called "data viewer" was developed. It consists of a table that displays a list of publications depending on the user's interaction with the knowledge map. If the user selects a node, it displays the author's publications. If an edge is selected, it displays the publications that a pair of authors have in common.

The knowledge map and data viewer always have their contents linked, since they exist to support each other. This means that the search and filter functionalities apply equally to both. These functionalities include filtering by the type of publication (such as books, articles, etc.) and a range for the date of publication. In addition to these filters, the user has the ability to execute a keyword search that generates a new knowledge map with publications that relate to that term. The system generates that map with a list of keywords that are similar to the user's query and returns both the map and the list. Finally, the user has the ability to discard unwanted keywords from that list, making the knowledge map more accurate to the user's specifications.

5.4. Overview

Figure 5 presents an overview of the main functionalities when applied to an author's page. The knowledge map and data viewer are the main elements of the page, followed by other statistics, such as the total count of publications (and more), bar charts for publications and projects in a given year, most common keywords and areas and top collaborators. Note that a similar page was developed for the institutional scope, as well as a page for managing the authors and importing their data from external sources.

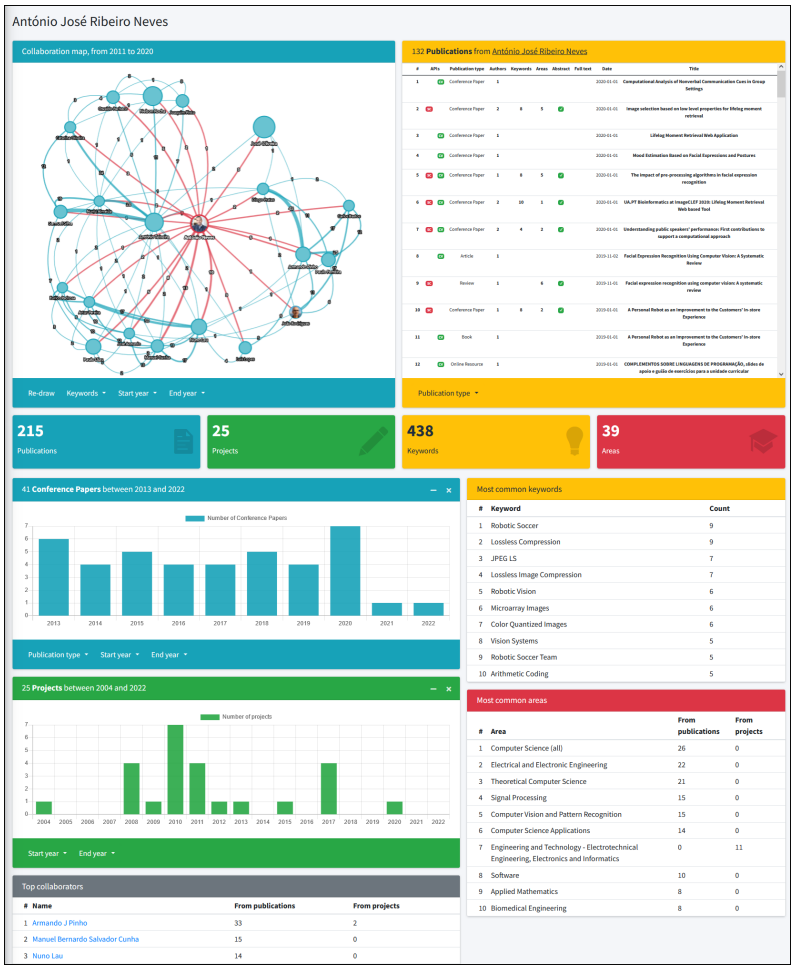

Figure 5. Overview of an author's page.

6. Conclusions

The proposed platform represents a unique way to enhance the management of scientific competences at an organization. It could become a key component for the administration of any academic institute.

With the proposed data collection module, it is possible to extract rich information from public APIs, respecting their limitations. Redundancy, or duplicate checking of the gathered information, was also a challenge that was solved based on the use of NLP techniques to find similar text across many publications and deem them equal or not. The knowledge extraction phase went through different visualization tools and their characteristics. Upon settling on one framework, it was shown that it is possible to represent raw data in an enhanced visual representation.

The data that were gathered were transformed into tables and charts, but the network graphs stood out as being a powerful interpretation of the current assets of an organization. Another tool, named "data viewer", was developed to improve the user experience when interacting with the knowledge map.

One feature that could be added is the ability to store citation information and enhance the existing tools with that added data. The system could also attempt to extract keywords from the publication's title, abstract and full-text content, making the keyword list more accurate. In its current state, this platform can run on any computer with *Python* and *Django* installed, however, it would be much easier to deploy if it was included in a *Docker* container, for example. This would help institutions to get the platform up and running much quicker, making it much more appealing.

Author Contributions: Conceptualization, J.G., A.T. and A.J.R.N.; methodology, J.G., A.T. and A.J.R.N.; software, J.G.; writing—original draft preparation, J.G.; writing—review and editing, J.G., A.T. and A.J.R.N.; supervision, A.T. and A.J.R.N. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the PRR - Recovery and Resilience Plan and by the NextGenerationEU funds at Universidade de Aveiro, through the scope of the Agenda for Business Innovation "NEXUS: Pacto de Inovação – Transição Verde e Digital para Transportes, Logística e Mobilidade" (Project no. 53 with the application C645112083-00000059).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 https://plos.org/ (accessed on 17 March 2023)
2 https://pubmed.ncbi.nlm.nih.gov/ (accessed on 17 March 2023)
3 https://www.elsevier.com/ (accessed on 17 March 2023)
4 https://www.cienciavitae.pt/?lang=en (accessed on 17 March 2023)
5 https://www.djangoproject.com/ (accessed on 17 March 2023)
6 https://www.python.org/ (accessed on 17 March 2023)
7 https://www.chartjs.org/ (accessed on 17 March 2023)
8 https://visjs.org/ (accessed on 17 March 2023)
9 https://github.com/ElsevierDev/elsapy (accessed on 17 March 2023)
10 https://www.scopus.com/ (accessed on 17 March 2023)
11 https://www.sciencedirect.com/ (accessed on 17 March 2023)
12 https://requests.readthedocs.io/en/latest/ (accessed on 17 March 2023)
13 https://www.nltk.org/ (accessed on 17 March 2023)
14 http://ieeta.pt (accessed on 17 March 2023)
15 https://github.com/joaogenio/knowledge-maps (accessed on 17 March 2023)

References

1. Gaviria-Marin, M.; Merigó, J.M.; Baier-Fuentes, H. Knowledge management: A global examination based on bibliometric analysis. *Technol. Forecast. Soc. Chang.* **2019**, *140*, 194–195. [CrossRef]
2. Kemp, D.Y. Knowledge Management in a Research & Development Environment—The Integration of Company Culture and Technology. Master's Thesis, Rochester Institute of Technology, Rochester, NY, USA, 2004.
3. Brazdil, P.; Trigo, L.; Cordeiro, J.; Sarmento, R.; Valizadeh, M. Affinity Mining of Documents Sets via Network Analysis, Keywords and Summaries. *LinguíStica InformáTica TraduçãO Mundos Que Cruzam Oslo Stud. Lang.* **2015**, *7*, 183–207. [CrossRef]
4. Iqbal, A.; Latif, F.; Marimon, F.; Sahibzada, U.F.; Hussain, S. From knowledge management to organizational performance: Modelling the mediating role of innovation and intellectual capital in higher education. *J. Enterp. Inf. Manag.* **2019**, *32*, 36–37. [CrossRef]
5. Thelwall, M. Bibliometrics to webometrics. *J. Inf. Sci.* **2008**, *34*, 605–621. [CrossRef]
6. Godin, B. On the origins of bibliometrics. *Scientometrics* **2006**, *68*, 109–133. [CrossRef]
7. Hirsch, J.E. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 741–754. [CrossRef] [PubMed]
8. Lee, J.H.; Segev, A. Knowledge maps for e-learning. *Comput. Educ.* **2012**, *59*, 353–364. [CrossRef]
9. Nesbit, J.C.; Adesope, O.O. Learning with concept and knowledge maps: A meta-analysis. *Rev. Educ. Res.* **2006**, *76*, 413–448. [CrossRef]
10. Xie, Z.; Ouyang, Z.; Li, J. A geometric graph model for coauthorship networks. *J. Inf.* **2016**, *10*, 299–311. [CrossRef]
11. Choobdar, S.; Ribeiro, P.; Bugla, S.; Silva, F. Comparison of Co-authorship Networks across Scientific Fields Using Motifs. In Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey, 26–28 August 2012; pp. 147–152. [CrossRef]
12. Cmap Software. Available online: https://cmap.ihmc.us (accessed on 17 March 2023).
13. Visual Understanding Environment. Available online: https://vue.tufts.edu (accessed on 17 March 2023).
14. Online Mind Mapping. Available online: https://www.mindmup.com (accessed on 17 March 2023).
15. Authenticus Portal. Available online: https://www.authenticus.pt/ (accessed on 17 March 2023).
16. Silva, F. Authenticus—Enabling the Identification and Validation of Portuguese Scientific Publications. In Proceedings of the euroCRIS Membership Meeting Autumn, Universidade do Porto, Porto, Portugal, 14–15 November 2013.
17. Domingues, F.A. Authenticus: Architecture and Mechanisms to Support a National Repository of Scientific Publications. Master's Thesis, Faculty of Sciences of the University of Porto, Porto, Portugal, 2015.
18. Open Knowledge Maps—A Visual Interface to the World's Scientific Knowledge. Available online: https://openknowledgemaps.org/ (accessed on 17 March 2023).
19. Kraker, P.; Kittel, C.; Enkhbayar, A. Open Knowledge Maps: Creating a Visual Interface to the World's Scientific Knowledge Based on Natural Language Processing. *J. Libr. Cult./Z. Bibl.* **2016**, *4*, 98–103. [CrossRef]
20. Pure—Leverage the world's leading Research Information Management System. Available online: https://www.elsevier.com/solutions/pure (accessed on 17 March 2023).
21. Interactive Scopus APIs. Available online: https://dev.elsevier.com/scopus.html (accessed on 17 March 2023).
22. Interactive ScienceDirect APIs. Available online: https://dev.elsevier.com/sciencedirect.html (accessed on 17 March 2023).
23. Ciência Vitae Swagger UI. Available online: https://api.cienciavitae.pt/docs/ (accessed on 17 March 2023).
24. Silva, J.M.B.; Silva, F. Feature Extraction for the Author Name Disambiguation Problem in a Bibliographic Database. In Proceedings of the Symposium on Applied Computing, SAC '17, Marrakech, Morocco, 3–7 April 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 783–789. [CrossRef]