



# **Automating Bibliometric Analysis of *CMU Portugal*: Uncovering Research Impact and Collaborative Networks**

**João Carlos Jerónimo Antunes**

Thesis to obtain the Master of Science Degree in

**Computer Science and Engineering**

Supervisors: Prof. Maria Inês Camarate de Campos Lynce de Faria  
Dr. Sílvia Manuela Azevedo de Castro

**Examination Committee**

Chairperson: Prof. Name of the Chairperson  
Supervisor: Prof. Maria Inês Camarate de Campos Lynce de Faria  
Member of the Committee: Prof. Luís Jorge Brás Monteiro Guerra e Silva

**June 2023**

**Declaration**

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

# Acknowledgments

I would like to thank my family for their friendship, encouragement, and caring over all these years, for always being there for me through thick and thin, and without whom this project would not be possible.

I would also like to acknowledge my dissertation supervisors, Prof. Maria Inês Camarate de Campos Lynce de Faria and Dr. Sílvia Manuela Azevedo de Castro for their insight, support, and sharing of knowledge that has made this thesis possible.

Additionally, I would like to thank João Fumega from the *Carnegie Mellon Portugal Program (CMU Portugal)* team for his insight and feedback.

A special thanks to Maria Inês Morais for caring and helping me through this process.

Last but not least, to all my friends and colleagues that helped me grow as a person and were always there for me during the good and bad times in my life. Thank you.

To each and every one of you – Thank you.



# Abstract

The *Carnegie Mellon Portugal Program (CMU Portugal)* program is an international collaboration between *Carnegie Mellon University (CMU)* and several Portuguese colleges, research institutes, and businesses. This program aims to put Portugal at the forefront of technological advancements by promoting education, research, innovation, and institutional network collaboration. To investigate and quantify the impact of *CMU Portugal's* initiatives in Portugal, one important component is to evaluate the quality of the resulting research output. Available bibliometric data from academic data repositories, like *Google Scholar*, is often used to rate academic research performance and researchers. Given this, it is possible to study the impact of *CMU Portugal's* initiatives by performing a bibliometric analysis of scientific publications by researchers under the scope of the *CMU Portugal* partnership. This master's thesis dissertation developed a platform that simplifies the online identification of research and academic output and uses factors, such as citation count and author's institution affiliations, to quantify the impact caused by *CMU Portugal*. This is to be implemented by extracting *CMU Portugal's* associated bibliometric data, from *Google Scholar* through the use of *Application Programming Interfaces (APIs)* and web scraping techniques. As such, an overview of the methodology used is provided throughout the document. To evaluate the usability of the final platform, we conducted interviews with users. We concluded that we were able to automate the process of extracting data from *Google Scholar* and had positive results regarding the platform's usability.

## Keywords

*Carnegie Mellon Portugal Program (CMU Portugal)*; International Partnership; Data Extraction; Bibliometric Data; *Google Scholar*; Research Impact.



# Resumo

O programa *Carnegie Mellon Portugal Program (CMU Portugal)* é uma colaboração internacional entre a *Carnegie Mellon University (CMU)* e várias faculdades, institutos de investigação e empresas portuguesas. Este programa tem como objectivo colocar Portugal na vanguarda dos avanços tecnológicos, promovendo a educação, a investigação, a inovação e a colaboração em rede institucional. Para investigar e quantificar o impacto das iniciativas da *CMU Portugal* em Portugal, um componente importante é avaliar a qualidade da produção de investigação resultante. Os dados bibliométricos disponíveis em repositórios de dados académicos, como o *Google Scholar*, são frequentemente utilizados para avaliar o desempenho da investigação académica e de investigadores. Assim, é possível estudar o impacto das iniciativas do *CMU Portugal* através de uma análise bibliométrica das publicações científicas dos investigadores da *CMU Portugal*. Esta dissertação de mestrado desenvolveu uma plataforma que simplifica a identificação online da investigação e da produção académica e utiliza factores, como o número de citações e a afiliação institucional dos autores, para quantificar o impacto causado pelo *CMU Portugal*. Este objectivo é implementado através da extracção de dados bibliométricos associados ao *CMU Portugal*, a partir do *Google Scholar*, através da utilização de *Application Programming Interfaces (APIs)* e de técnicas de web scraping. Como tal, ao longo do documento é apresentada uma visão geral da metodologia utilizada. Para avaliar a usabilidade da plataforma final, realizámos entrevistas com utilizadores. Concluímos que conseguimos automatizar o processo de extracção de dados do *Google Scholar* e obtivemos resultados positivos relativamente à usabilidade da plataforma.

## Palavras Chave

*CMU Portugal*; Colaboração Internacional; Extracção de Dados; Dados Bibliométricos; *Google Scholar*; Impacto da Investigação.





# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Problem Definition . . . . .	1
1.2	Objective . . . . .	2
1.3	Proposed Solution . . . . .	3
1.4	Contributions . . . . .	3
1.5	Document Organization . . . . .	4
<b>2</b>	<b>About CMU Portugal</b>	<b>7</b>
2.1	Talent Development . . . . .	9
2.2	Knowledge Creation . . . . .	9
2.3	Innovation and Entrepreneurship . . . . .	10
2.4	Communication and Outreach . . . . .	10
<b>3</b>	<b>Related Work</b>	<b>11</b>
3.1	International Portuguese Partnerships . . . . .	12
3.1.1	<i>University of Texas Austin Portugal (UT Austin Portugal)</i> . . . . .	12
3.1.2	<i>MIT Portugal Program (MPP)</i> . . . . .	13
3.2	International Collaboration Impact Assessment . . . . .	14
3.2.1	<i>MPP's Mobility Program</i> . . . . .	14
3.2.2	<i>CMU Portugal's Faculty Exchange Program</i> . . . . .	15
3.2.3	<i>MPP's Difference-in-difference Network Formation and Research Re-orientation</i> . . . . .	15
3.3	Scientific and Academic Research Data Repositories . . . . .	16
3.3.1	<i>Google Scholar</i> . . . . .	16
3.3.2	<i>Web of Science (WoS)</i> . . . . .	17
3.3.3	<i>Scopus</i> . . . . .	17
3.3.4	<i>Open Researcher and Contributor ID (ORCID)</i> . . . . .	17
3.4	Research Output Impact Assessment . . . . .	17
3.4.1	Ranking by Relevance and Citation Counts . . . . .	18
3.4.2	Tracking Scholarly Output through <i>Google Scholar</i> . . . . .	18

3.4.3	<i>Google Scholar</i> Evaluation of Journal Articles Impact in Education . . . . .	19
3.5	<i>Application Programming Interfaces (APIs)</i> Description . . . . .	19
3.5.1	<i>SerpAPI</i> . . . . .	20
3.5.2	<i>ScraperAPI</i> . . . . .	21
3.5.3	<i>scrapy</i> and <i>scholarly</i> libraries . . . . .	22
<b>4</b>	<b>Implementation</b>	<b>23</b>
4.1	Requirements' Analysis . . . . .	24
4.1.1	Authors . . . . .	24
4.1.2	Publications . . . . .	25
4.1.3	Platform . . . . .	26
4.2	Available Information . . . . .	26
4.2.1	Current <i>Excel</i> Data Organization . . . . .	26
4.2.2	<i>Google Scholar's</i> Data . . . . .	27
4.3	Tools and Technology . . . . .	30
4.3.1	Tools Selection . . . . .	31
4.3.2	System Dependencies . . . . .	34
4.4	System Architecture . . . . .	35
4.5	Methodology . . . . .	36
4.5.1	Data Scraping . . . . .	36
4.5.2	Data Update . . . . .	40
4.5.3	Data Crossing . . . . .	44
4.5.4	Final Platform . . . . .	53
4.5.5	External Features . . . . .	60
<b>5</b>	<b>Evaluation</b>	<b>63</b>
5.1	User Characterization . . . . .	64
5.2	Platform Validation . . . . .	65
5.3	Tasks Execution Results . . . . .	67
5.4	Usability Questionnaire Results . . . . .	68
5.5	Results Discussion . . . . .	70
<b>6</b>	<b>Conclusion</b>	<b>73</b>
6.1	Final Discussion . . . . .	73
6.2	Current Limitations . . . . .	74
6.2.1	Author's Exclusion . . . . .	75
6.2.2	Publications Outside the Scope of <i>Carnegie Mellon Portugal Program (CMU Portugal)</i>	75
6.2.3	<i>Google Scholar's</i> Data Update . . . . .	75

6.2.4	International and Inter-institutional Collaboration Identification: . . . . .	76
6.2.5	External Features . . . . .	77
6.2.6	<i>JavaScript Object Notation (JSON)</i> Data Handling . . . . .	77
6.3	Future Work . . . . .	77
6.3.1	<i>ORCID</i> Profiles . . . . .	77
6.3.2	Documents Affiliation Identification . . . . .	78
6.3.3	Use <i>CMU Portugal</i> as Keyword for Searching . . . . .	78
6.3.4	Create Database for Extracted Data . . . . .	79
6.3.5	Correct Usability Issues . . . . .	79
<b>Bibliography</b>		<b>79</b>
<b>A Demographic Questionnaire</b>		<b>85</b>
A.1	Preparation . . . . .	85
A.2	Questionnaire . . . . .	85
<b>B User Guide</b>		<b>89</b>
B.1	Introduction . . . . .	89
B.2	Evaluation . . . . .	90
<b>C Usability Questionnaire</b>		<b>93</b>
C.1	Final Balance . . . . .	93



# List of Figures

2.1	<i>CMU Portugal Conceptual Map</i>	8
4.1	<i>PhD. and affiliate students with a Google Scholar Profile</i>	27
4.2	<i>Author's Google Scholar Profile Page</i>	28
4.3	<i>Publication's Google Scholar Profile Page</i>	29
4.4	<i>Conference Paper's Google Scholar Profile Page</i>	30
4.5	<i>Information Gathering Architecture</i>	35
4.6	<i>"links.json" Publication Entry</i>	37
4.7	<i>"links.json" Publication Order</i>	39
4.8	<i>"barsTemp.json" Publication Entry</i>	40
4.9	<i>"deletedPublications.json" File Structure</i>	42
4.10	<i>"authorsInfoList.json" Author Entry</i>	43
4.11	<i>Citations Error Example</i>	45
4.12	<i>Citations' Data</i>	46
4.13	<i>"authorsData.js" "CMUPortugal" Field</i>	50
4.14	<i>"authorsData.js" "authors" Field</i>	51
4.15	<i>"authorsData.js" "publications" Field</i>	52
4.16	<i>"authorsDataExcel.xlsx" File</i>	53
4.17	<i>Global Dashboard's Quantitative Metrics</i>	54
4.18	<i>Global Dashboard's Charts</i>	55
4.19	<i>Author's List Page</i>	56
4.20	<i>Author's Profile Card and Metrics</i>	57
4.21	<i>Author's Profile Publications List</i>	57
4.22	<i>Author's Profile Collaboration Students List</i>	58
4.23	<i>Publications List Table</i>	59
4.24	<i>Publications' Profile Page</i>	59
4.25	<i>Modules' Code Graphical Interface</i>	61

4.26	<i>Delete Publications Graphical Interface</i>	61
5.1	<i>User's Experience Circular Chart</i>	64
5.2	<i>User's Easiness Circular Chart</i>	65
5.3	<i>Heuristic Evaluation Results</i>	70
6.1	<i>Author's ORCID profile page</i>	78

# List of Tables

4.1	<i>APIs</i> Tools Selection . . . . .	32
4.2	Platform Tools Selection . . . . .	33
5.1	Recorded Tasks' Times . . . . .	67
5.2	Heuristics Results . . . . .	68
C.1	Heuristics Rating . . . . .	94





# Acronyms

<b>CMU</b>	Carnegie Mellon University
<b>CMU Portugal</b>	Carnegie Mellon Portugal Program
<b>CRUP</b>	Conselho de Reitores das Universidades Portuguesas
<b>ERI</b>	Entrepreneurial Research Initiative
<b>ERP</b>	Exploratory Research Project
<b>FCT</b>	Portuguese Foundation for Science and Technology
<b>ICT</b>	Information and Communication Technologies
<b>LSCR</b>	Large-Scale Collaborative Research
<b>API</b>	Application Programming Interface
<b>UT Austin Portugal</b>	University of Texas Austin Portugal
<b>UT Austin</b>	University of Texas Austin
<b>MPP</b>	MIT Portugal Program
<b>MIT</b>	Massachusetts Institute of Technology
<b>WoS</b>	Web of Science
<b>h-index</b>	Hirsch Index
<b>WWW</b>	World Wide Web
<b>HTML</b>	Hypertext Markup Language
<b>HTML5</b>	Hypertext Markup Language
<b>Ph.D.</b>	Professional Doctorate Degree
<b>PT</b>	Portugal
<b>DOI</b>	Digital Document Identifier
<b>PDF</b>	Portable Document Format
<b>CSS</b>	Cascading Style Sheets

<b>Power BI</b>	Power Business Intelligence
<b>D3</b>	Data-Driven Documents
<b>JSON</b>	JavaScript Object Notation
<b>HTTP</b>	Hypertext Transfer Protocol
<b>CORS</b>	Cross-Origin Resource Sharing
<b>ORCID</b>	Open Researcher and Contributor ID
<b>NLP</b>	Natural Language Processing
<b>NER</b>	Named Entity Recognition
<b>WWW</b>	World Wide Web
<b>ORCID</b>	Open Researcher and Contributor ID
<b>SERP</b>	Search Engine Results Page
<b>URL</b>	Uniform Resource Locator
<b>CSV</b>	Comma-Separated Values
<b>IP</b>	Internet Protocol





# 1

## Introduction

### Contents

1.1 Problem Definition . . . . .	1
1.2 Objective . . . . .	2
1.3 Proposed Solution . . . . .	3
1.4 Contributions . . . . .	3
1.5 Document Organization . . . . .	4

### 1.1 Problem Definition

Since 2006, there have been several international research and innovation collaborations between institutions in the United States and Portuguese organizations and universities. One of these collaborations is *Carnegie Mellon Portugal Program (CMU Portugal)* [1] and one of its main objectives is to create a time-lasting impact and influence over scientific and academic research and education, as well as to promote collaboration networks across several research institutions in Portugal [2].

When analyzing to what extent *CMU Portugal* has had an impact on international scientific and academic research, a key aspect is assessing the quality of the resulting research output. Several studies

( [3] [4] [5] [6] [7] [8] [9] ) have proceeded on how to quantify this impact and to attribute factors, measures, and criteria to evaluate the quality and importance of publications as well as the individual contribution of researchers/authors. Some of these studies used bibliometric data to analyze and evaluate academic content and its authors.

Bibliometric information can vary from listings of publications and authors to the number of citations and linked institutions. As such, this information is significant because the number of other papers that have cited the publication or author can be calculated and reflect the importance of publications or writers [6].

*CMU Portugal's* bibliometric data on outcomes can provide factors to quantify the program's impact on academic research. However, to gather these factors, it is essential to keep track of this data. Furthermore, the list of linked institutions and organizations that have participated in *CMU Portugal's* initiatives demonstrates that the program has promoted worldwide research collaboration.

## 1.2 Objective

As a result, our main objective is to make the process of tracking bibliometric data concerning *CMU Portugal's* research output easier and more automated. To accomplish this, we must first determine which documents and researchers fall under the scope of this program. Another one of our primary objectives in this project is to analyze *CMU Portugal's* bibliometric data. Bibliometrics is a quantitative analysis of publications, citations, and other bibliographic data to assess the impact and productivity of academic research.

By conducting an analysis of *CMU Portugal's* bibliometric data, we aim to gain valuable insights into the research output and impact of the organization. One aspect of our analysis involves evaluating the productivity of *CMU Portugal's* authors and their research outcomes. By examining publication counts, citation counts, and publication trends over time, we can identify highly productive authors and assess the impact of their research contributions.

Furthermore, we aim to analyze the collaboration networks within *CMU Portugal*. By studying co-authorship patterns and identifying international collaborations, we can understand the extent of knowledge exchange and partnerships facilitated by *CMU Portugal*.

Another goal, once we have gathered the needed data, is to cross-reference the information and visualize this data on an interactive platform. This way, we can measure the impact caused by *CMU Portugal* by evaluating the displayed bibliometric data.

## 1.3 Proposed Solution

Our proposed solution focuses on extracting data for authors who are *Professional Doctorate Degree (Ph.D.)* students or affiliated with *CMU Portugal*. This group of authors was chosen because, outside this scope, authors did not have *Google Scholar* profiles, also, each student has a *Carnegie Mellon University (CMU)* and *Portugal (PT)* advisor, where each advisor represents its respective institution. Because of this, we can identify international publications by verifying if both of these advisors are credited as authors in the publication. Additionally, we limited the scope of publications to those published within the timeframe when the students were associated with *CMU Portugal*, considering their start and end research years, with an additional one-year margin. This approach ensures that the extracted data represent the research activities during the affiliation period.

To implement the proposed solution, we employed web scraping techniques to scrape data directly from the *Google Scholar*. For authors, we extracted details such as names, affiliations, and citation counts. We also collected information about the publications, including titles, publication dates, and citation counts. This data provides insights into the research output, impact, and overall contribution of *CMU Portugal*.

To enhance the accessibility and usability of the collected data, we developed a dashboard that visualizes the extracted information. This dashboard serves as a platform for users to explore and interact with the data in an intuitive and user-friendly manner. Through various charts, graphs, and tables, users can gain insights into publication trends, author profiles, collaboration networks, and research impact.

By combining the data extraction methodology and the development of the information dashboard, our proposed solution provides *CMU Portugal* with an efficient and effective means of extracting, organizing, and visualizing data from *Google Scholar*. This solution empowers *CMU Portugal* to evaluate the impact of its research programs, authors, and publications.

## 1.4 Contributions

The automation of data extraction from *Google Scholar*, coupled with the development of an information dashboard to visualize the collected data, brings several significant contributions to *CMU Portugal* in terms of data management, updating, and visualization. These contributions have the potential to enhance *CMU Portugal's* ability to stay updated on the research activities and collaborations within its network.

- **Efficient Data Extraction:** The automated extraction process significantly reduces the manual effort required to collect data from *Google Scholar*. By leveraging web scraping techniques and appropriate algorithms, we can retrieve relevant information about *CMU Portugal's* authors and

publications in a more efficient and timely manner. This ensures that the dashboard's data remains up-to-date and reflective of the latest research outputs and collaborations.

- **Comprehensive Author and Publication Information:** The automation of data extraction allows for a comprehensive collection of author and publication information. By systematically retrieving details such as author affiliations, publication titles, co-authors, citation counts, and publication dates, the dashboard provides a holistic view of the research output and impact of *CMU Portugal's* program. This comprehensive information enables stakeholders to gain insights into the productivity and influence of *CMU Portugal's* researchers.
- **Real-time Data Visualization:** The development of an information dashboard provides a user-friendly and visually appealing interface to explore and analyze the collected data. The dashboard's visualization capabilities enable stakeholders to understand the research landscape. By visualizing data on publications, authors, collaborations, and impact indicators, *CMU Portugal* can better monitor its progress and assess the effectiveness of its programs and initiatives.
- **Facilitated Data Updating:** The automated data extraction process, coupled with the information dashboard, streamlines the data updating process. With the ability to extract data on-demand, *CMU Portugal* can easily refresh the dashboard with the latest information.

## 1.5 Document Organization

This thesis is structured as follows: First, *Section 2* gives an introduction to what the *CMU Portugal* program consists of, as well as the collaboration's main goals and approaches to achieve them. Then, *Section 3* discusses related work, whereas in *Section 3.1* it explores other international partnerships with Portuguese institutions. *Section 3.2* goes through several studies that are directed towards these international partnerships and analyzes how much impact each of the programs' initiatives and projects has caused. In addition, *Section 3.3* explains what academic research data repositories are while enumerating three of them. In *Section 3.4* we explore other studies that have performed an analysis of how data repositories can be used as a tool to access scientific impact from research output and individual researchers. *Section 3.4* additionally describes *Google Scholar's* algorithm that ranks its pages and documents. The next section is *Section 3.5* where we give an introduction to scraping tools that can be used to extract data from website pages. The proposed solution and its implementation are presented in *Section 4*. In *Section 4.1* we describe the requirements and the information that we wish to extract from *Google Scholar*. *Section 4.2*, describes which information is available to be extracted and which information we have access to. Regarding *Section 4.3*, we explain which tools were used in our implementation and why they were chosen. Additionally, *Section 4.4* describes the architecture for our



implementation and its components. Followed by this, is *Section 4.5* where we explain the methodology used to extract data from *Google Scholar* and the methodology that developed the final platform. Next is *Section 5* where we describe the evaluation process that was used to evaluate the final platform and the results from this evaluation. Followed by this is *Section 6*, which presents a summary of this thesis's most important key features, current limitations of our implementation, and future work.



# 2

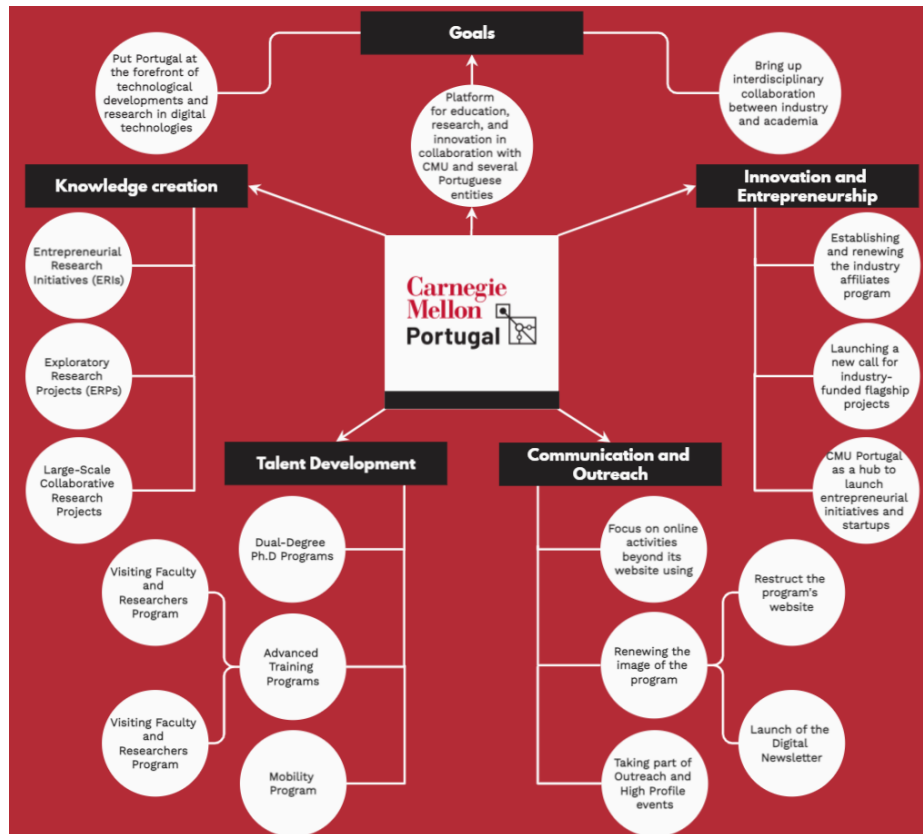
## About CMU Portugal

### Contents

2.1 Talent Development . . . . .	9
2.2 Knowledge Creation . . . . .	9
2.3 Innovation and Entrepreneurship . . . . .	10
2.4 Communication and Outreach . . . . .	10

According to the *CMU Portugal* 2018/2019 annual report [2], the *CMU Portugal* is an international platform for education, research, and innovation that was founded in 2006. This program includes collaboration with *CMU* and several Portuguese universities, research institutions, and companies.

This initiative aims to put Portugal at the forefront of technological developments and research in digital technologies and the area of *Information and Communication Technologies (ICT)* to encourage and promote cutting-edge research and world-class graduate education. This project is supported by the *Portuguese Foundation for Science and Technology (FCT)*, sponsored by the *Conselho de Reitores das Universidades Portuguesas (CRUP)*, and co-financed by *CMU* and industry partners. Currently, *CMU Portugal* supports 12 *Large-Scale Collaborative Research (LSCR)* projects led by Portuguese companies with partnerships with Portuguese universities/research institutions and *CMU*. In *Figure 2.1* we can find a conceptual map for *CMU Portugal*'s current activities.



**Figure 2.1: CMU Portugal Conceptual Map**

Since 2006, *CMU Portugal* has been through two previous phases (Phase 1 and Phase 2). During these phases, *CMU Portugal* successfully promoted the development of talent and internationalization of Portuguese universities and encouraged cooperation between universities, independent researchers, and Portuguese companies. Currently, in their 3rd phase, which began in 2018 and is expected to last until 2030, the major objective is “to bring up interdisciplinary collaboration between industry and academia across different levels of “big data” development stack” [2].

The current collaborative network of *CMU Portugal* consists of several Portuguese universities represented by *CRUP*, Associate Laboratories in the area of *ICT*, other Portuguese research institutions, 10 *CMU* Departments, almost 150 companies and over 400 faculty, senior researchers at both Portugal and *CMU*, collaboration agreements with 14 new industrial affiliates, and *ICT* leaders in Portugal and globally.

To achieve the mentioned goals, *CMU Portugal* has the following initiatives and programs that are listed in the *CMU Portugal* 2018/2019 Annual Report [2]:

- **Talent Development**
- **Knowledge Creation**

- ***Innovation and Entrepreneurship***

- ***Communication and Outreach***

Each initiative will be described in the following sections:

## 2.1 Talent Development

Following the *CMU* Program, Portuguese universities and *CMU* offer *Dual-Degree Doctoral Programs* in several areas in which successful candidates are awarded two *Ph.D.* degrees, where each one is, respectively, from *CMU* and one of the Portuguese universities of this program. *CMU Portugal* also features its *Mobility Program* which contains the *Visiting Faculty and Researchers* and the *Visiting Students programs*. The *Visiting Faculty and Researchers* program has existed since Phase 1 and is directed toward Post-Doctoral researchers and encourages the integration of faculty from Portuguese universities into international knowledge networks. The *Visiting Students* program offers master's students the opportunity to participate in a research project at *CMU*. Adding to this initiative, Portuguese universities, *CMU* departments, and industry partners intend to establish a *Advanced Training Programs* in the areas of *Data Science* and *Machine Learning* and *User Experience Design*.

## 2.2 Knowledge Creation

With this initiative, *CMU Portugal* program intends to launch *Small Seed Funding Research* projects to create *Small-Scale Research* collaborations. This includes the *Entrepreneurial Research Initiatives (ERIs)* and *Exploratory Research Projects (ERPs)*, and the involvement in *Large-Scale Collaborative Research* projects. The *ERIs* program consists of science, engineering, management, and policy projects that merge research, innovation, and advanced training initiatives in collaboration with several companies. To manage and monitor these projects, there is a group of researchers from two Portuguese universities, one from *CMU*, and at least one corporate partner. Regarding the *ERPs*, these aim to foster new initiatives and promote information and communication technologies projects and integrative research in strategic emerging areas. As mentioned before, *CMU Portugal* is also involved in several *Large-Scale Collaborative Research* projects to include industrial research and experimental development activities and to create new products, services, processes, and systems.

## 2.3 Innovation and Entrepreneurship

*CMU Portugal* has also been at the center for faculty members, students, and alumni to launch their entrepreneurial initiatives and startups. To follow this concept, *CMU Portugal* aims to establish and renew the *Industry Affiliates Program* by developing close relationships with Portuguese industry and ties between companies that are members of the *Industrial Affiliates Program*, as well as launch a new call for industry-funded flagship projects. However, large-scale and *ERIs* projects can also be co-financed and supported by non-member companies through the *Industrial Affiliates Program*.

## 2.4 Communication and Outreach

This particular initiative focuses on renewing the image of the program as a platform for international scientific collaboration. This consists of restructuring the *CMU Portugal's* website by rethinking its navigation, usage, and cross-platform support, and launching a digital newsletter. Another goal is to focus on online activities beyond its website and extend the information flow through social media networks and the press to highlight the program's research outputs and faculty, students, and alumni achievements. It is also intended to target e-mail messages to specific audiences in the scientific and research communities. *CMU Portugal* has also been part of outreach events to interact with a broader audience while promoting the program and organizing high-profile events to reach out to strategic stakeholders and entities.

# 3

## Related Work

### Contents

---

3.1 International Portuguese Partnerships . . . . .	12
3.2 International Collaboration Impact Assessment . . . . .	14
3.3 Scientific and Academic Research Data Repositories . . . . .	16
3.4 Research Output Impact Assessment . . . . .	17
3.5 <i>Application Programming Interfaces (APIs) Description</i> . . . . .	19

---

In addition to *CMU Portugal*, there are various other partnerships and collaborations around the world that strive to promote academic research and foster international collaboration. These initiatives recognize the importance of knowledge exchange and the benefits of interdisciplinary cooperation in advancing scientific discoveries.

To facilitate the exploration and evaluation of academic research, there are data repositories dedicated to storing scholarly output. These repositories serve as databases that provide a platform for researchers to showcase their work and make it accessible to the wider academic community. Within these repositories, users are often assigned unique profiles that aggregate their publications, citations, and other relevant bibliographic information. One of the key advantages of these data repositories is their ability to generate bibliometric data, which quantifies the impact and reach of academic research.

By examining citation counts, *Hirsch Index (h-index)*, and other bibliometric indicators, researchers can gain insights into the significance and reception of their research outputs [9].

To extract this academic data from the repositories, various tools and techniques enable users to harvest information from the repositories' web pages, allowing for analysis and evaluation of research impact.

In the following sections, we describe in more detail the topics above.

### 3.1 International Portuguese Partnerships

In the *Setting-up an international science partnership program: a case study between Portuguese and US research universities* publication [1] it is performed an analysis on the background and logic behind the early development and structuring of three ongoing international partnerships between Portugal and US universities:

- ***University of Texas Austin Portugal (UT Austin Portugal)***
- ***MIT Portugal Program (MPP)***
- ***CMU Portugal***

This essay concludes that these three worldwide collaborations can operate as change agents, without financial support being the most essential element affecting university participation. Instead, contributions to international research, institution modernization efforts, and faculty mobility programs serve as important catalysts for policymakers to learn from other international partnerships and develop the long-term evolution and involvement of Portuguese universities, institutions, and researchers with world-renowned US universities.

Following in the previous introduction to *CMU Portugal* and this last note, these referred partnership programs share the same goal as *CMU* on being agents of change and have created initiatives to support interdisciplinary and international collaboration with Portuguese institutions:

#### 3.1.1 *UT Austin Portugal*

**A – Overview** According to its website [10], the *UT Austin Portugal* program had its start in 2007, and it is a partnership program in Science and Technology between the *University of Texas Austin (UT Austin)* and the *FCT*. This program has the support of the *Ministry of Science, Technology and Higher Education* in close partnership with *CRUP*.



**B – Goal** The goal of this partnership is to promote the development and new frontiers of knowledge in worldwide emerging research themes by fostering the evolution of higher education, research, and commercialization activities, as well as boost the engagement between *UT Austin* and Portuguese scientists and companies at a large scale in international research, technology transfer, and commercialization activities [10].

**C – Innovation** This initiative continues to help bring cutting-edge Portuguese technologies to international markets by investing in differentiated, advanced, and modern training. Also, it fosters entrepreneurial initiatives to encourage the development and evolution of new technology directed toward international markets. This allows professional guidance and mentoring on a corporate level and professional training that leads to network growth, the acceleration and launch of Portuguese start-up companies, and the generation of new employment and market opportunities [10].

### 3.1.2 *MPP*

**A – Overview:** Created in 2006 the *MPP* is a strategic international collaboration between the *Massachusetts Institute of Technology (MIT)* and Portuguese research institutions and universities, with partnerships with the Portuguese government and industry partners. “The *MPP* was the first and the largest of the Portuguese collaborations” [3]. The *MPP*’s current goal is to establish an interactive, and long-lasting collaborative platform that deals with the major challenges of global and social impact in the following research areas: “Climate Science and Climate Change, Earth Systems: Oceans to Near Space, Digital Transformation in Manufacturing, and Sustainable Cities” [11].

**B – Goal** The *MPP* website [11] describes the goal of the program as the promotion and encouragement of collaborative research between *MIT* and Portuguese universities, companies, research institutes, and laboratories to create innovative, high-impact ideas for research projects directed to current and complex challenges that our society faces. An output of this goal is to help Portugal have a global impact by reinforcing the Portuguese academic and industrial ecosystem by creating technological solutions and applied research.

**C – Innovation** The *MPP* has established connections between *MIT* and Portuguese industry and engineering by developing initiatives like the *Building Global Innovators accelerator*, *E3 Forum* and the *International Workshop of Innovating (IWI)*. The *MPP* aims to continue fostering innovation in all fields of study, including the recruitment of numerous professors of innovation and entrepreneurship from Portuguese universities [11].

All the previously referred programs aim to bring up international collaboration in advanced research fields in emerging worldwide areas, as well as to strengthen Portugal's "knowledge base and international competitiveness through a strategic investment" [5]. These initiatives also seek to demonstrate that investing in research and developing more innovative methods for higher education may be a long-term answer for the economy and improve the current Portuguese education system's quality.

However, to conclude this, it is needed to create several factors and criteria on which we can evaluate the overall impact caused by the collaboration programs and verify if these can lead to lasting changes in Portuguese research, industry, companies, laboratories, universities, and laboratories, therefore, bringing Portugal to the forefront of technological development and global competition.

In *Section 3.2, 3.3, and 3.4* we can find several articles that have proceeded with exploring factors, parameters, and data to assess this impact.

## 3.2 International Collaboration Impact Assessment

### 3.2.1 *MPP's* Mobility Program

The *Seeding Change through International University Partnerships: The MIT-Portugal Program as a Driver of Internationalization, Networking, and Innovation* document [3] analyses the impact of the *MPP* by assessing the visibility and attractiveness of foreign students to Portuguese universities and international exposure since the start of this program. These factors can lead to international research networking, and according to this essay, international exposure has been fostered by the *MPP's* mobility programs, and since the start of the *MPP*, the number of international applications has increased. Another analyzed component was networking and clustering among universities, since "for students, the network is a central ingredient of their education" [3]. It is also worth analyzing the fact that this initiative also promoted direct ties between research activities and industrial stakeholders.

This study used three key items as metrics in order to verify if the *MPP* has caused considerable change:

1. **Internationalization:** This element allows "the cross-border flow of skills and knowledge and research productivity" [3] between countries and institutions.
2. **Selectivity:** In order to assess if highly skilled students were being attracted to Portuguese universities and institutions and if Portugal was being a target of international exposure.
3. **Clustering and Critical Mass Building:** This element allowed this study to evaluate the level of networking between institutions and universities.

4. **Innovation Orientation:** At what level and how was knowledge shared between faculty and students, and how would creating co-teaching courses and visiting-faculty programs lead to innovation in lecturing and entrepreneurship.

This study concludes that international collaborations can serve as forces for systemic change by demonstrating that *MPP* has had a considerable influence on Portuguese institutions. Additionally, it is said that Portugal's cooperative strategy serves as a blueprint for creating a focused base of human capital, research, and innovation suited for sustained economic growth.

### 3.2.2 *CMU Portugal's Faculty Exchange Program*

The *Faculty-exchange programs promoting change: motivations, experiences, and influence of participants in the Carnegie Mellon University-Portugal Faculty Exchange Program* [4] document explores the *CMU Portugal* faculty exchange program and how this "access to new academic environments, ideas, students, and colleagues" [4] can lead to the improvement of the teaching, research, and network quality. For this purpose, several research questions were created:

- What motivated faculty members to be part of the *CMU Portugal's* faculty exchange program?
- What contributions to new teaching and research understanding were generated by the program?
- What were the effects the program had on the faculty members on an individual and organizational level?

These questions were answered by collecting feedback from the participants of the *CMU Portugal* Faculty exchange program. At the end of this document, the authors conclude that the collected answers to the listed features above provide "insights for policymakers seeking to implement faculty-mobility programs in the future" [4].

### 3.2.3 *MPP's Difference-in-difference Network Formation and Research Re-orientation*

It is also important to consider that to evaluate the overall impact of a research collaboration initiative, it is necessary and imperative to gather data and information on the collaboration's scientific output and to define metrics for this data.

The *How complex international partnerships shape domestic research clusters: Difference-in-difference network formation and research re-orientation in the MIT Portugal Program* article [5] proposes a combination between bibliometric network analysis, difference-in-difference program evaluation, statistical matching techniques, and system architecture analysis to study this impact while taking into consideration four major current policy trends: *University-centrism*, *Collaboration*, *Internalization* and *Growing structural complexity*. We can assess these trends by analyzing the following parameters and goals:

1. **The high impact of research:** We evaluate this by analyzing the research output, the average impact factor per publication, which translates to the quality of each publication, the average number of publications per year, the number of publishing years, the number of authors, and the average number of citations per publications, which in this case translates to the overall visibility of the article.
2. **Empowerment of new generations of scientists:** This effort is investigated by comparing two sub-groups, those with publishing records of less than four years when the program began and those with more than 15 years' worth of publications.
3. **Encouragement of Portuguese collaboration with *MIT*:** This initiative is evaluated by analyzing the number of collaboration links between Portuguese institutions and *MIT*.
4. **Shaping research directions:** This parameter is assessed by observing the evolution of publication activity of *MPP-affiliated* faculty in terms of their content and by analyzing shifts in publication patterns.

This study concludes that international partnerships, the establishment of clusters, and the reorientation of research can have a substantial impact on the "hosting" country. Additionally, they contend that their approach offers a useful tool for assessing complex capacity-building programs and initiatives.

### 3.3 Scientific and Academic Research Data Repositories

The article entitled *The use of bibliometrics to measure research performance in education sciences* [7] article investigates the performance and impact of educational research professors through the use of bibliometric data from *Google Scholar* and *Web of Science (WoS)* platforms while doing a comparison between the two. This study [7] concludes that the bibliometric data from *Google Scholar* and *WoS* does reflect a correct impact evaluation of scholars with good research performance. Another conclusion is that there is a good balance trade-off between output quantity and outcome quality, which translates to, respectively, the number of publications and citation counts.

#### 3.3.1 *Google Scholar*

*Google Scholar* is a data repository directed towards scientific and academic output. This platform makes searching for relevant work across several fields of scientific research and literature a simpler process. It features a search engine for peer-reviewed journal articles, theses, conference papers, books, and book chapters, and its search results include an ordered list of publications' titles, authors, year, source information, redirecting links to full-text documents, citation count, a list of citing documents,

and hyperlinks to these documents [6]. This platform features a *PageRank* algorithm to sort the included publications [6]. This algorithm is explained in more detail in *Section 3.4.3*.

### **3.3.2 WoS**

*WoS* is a web-based research platform that offers a thorough and diverse library of bibliographic and citation data from academic publications, conference proceedings, and other sources. These publications are classified by field of research, country, and language [12]. In addition, *WoS* is considered to be the “world’s most trusted publisher-independent global citation database” [13].

### **3.3.3 Scopus**

*Scopus* is a peer-reviewed literature database that contains abstracts and citations for books, journals, and conference proceedings. The research output in the areas of science, technology, medicine, social sciences, and the arts and humanities is thoroughly analyzed by *Scopus* [14]. This scientific journal search and indexing database classifies its publications by an array of fields and filters, such as author searches, the field of research, citation indexes, country, and language [12].

### **3.3.4 Open Researcher and Contributor ID (ORCID)**

*ORCID*, is a non-profit organization that provides a unique digital identifier for researchers and scholars. The primary purpose of *ORCID* is to address the challenge of disambiguation of authors and contributors in scholarly communication. By assigning a persistent and unique identifier to each researcher, *ORCID* enables accurate and reliable attribution of their work. One of the features of *ORCID* is its ability to store and display information about a researcher’s affiliations over the years. Researchers can associate their profiles with various institutions, such as universities, research organizations, or industry affiliations, and indicate the time periods during which they were affiliated with each institution. This allows for a view of an individual’s academic journey and the collaborations and affiliations they have been a part of throughout their career [15].

## **3.4 Research Output Impact Assessment**

The last section mentioned possible individual factors, such as the number of publications or citation count, that could be used as a benchmark for research contribution assessment. The following studies discuss how to measure the impact of scientific publications. Even though these studies are not directed toward an international collaboration platform such as the ones referenced before, they explore how

scientific and academic data repository platforms and available features can reflect the research impact of scientific output and the respective contributions of each author/researcher.

### 3.4.1 Ranking by Relevance and Citation Counts

The conducted study in the *Ranking by relevance and citation counts, a comparative study: Google Scholar, Microsoft Academic, WoS, and Scopus* [8] article performs a comparison analysis between *Google Scholar*, *Microsoft Academic*, *WoS* and *Scopus* while asking how the methods and ranking algorithms featured in these data repositories “increase the visibility of, and the number of visits to, a web page through its ranking on the search engine results pages” [8].

The conclusion of this document, [8] states that both *Google Scholar* and *Microsoft Academic* rely mostly upon their citation count ranking algorithms. On the other hand, the *Scopus* search engine does not take into consideration the publication’s citation count in its ranking process.

Lastly, the *WoS* platform performed two different ranking algorithms on two different data collections: In the first data collection, the number of citations was not considered in the ranking system. Instead, it only considered the position and frequency of keywords, in the second data collection, oddly, the ranking algorithm was almost entirely based on citation count.

### 3.4.2 Tracking Scholarly Output through Google Scholar

In the *Using Google Scholar to track the scholarly output of research groups* publication [9], is performed a study on how to demonstrate the scholarly output of a research program over time. This study was conducted by creating *Google Scholar* profiles for five different research groups and analyzing how the automatically generated scholarly output and citation counts of individual researchers reflect the influence and impact of each research group.

The Researcher’s profile page from *Google Scholar* “provides a method to demonstrate the impact of a research program over time both within and beyond institutions” [9] since the *Google Scholar* platform tracks automatically the citation counts and the scholarly output of individual researchers. According to the *Using Google Scholar to estimate the impact of journal articles in education* document [6], this profile makes it possible to rank authors according to their citation count and the *h-index* in which the first *h* articles from the author’s document list, sorted according to citations number, all have at least *h* citations, and the remaining articles all have less than *h* citations.

Researchers can thus have a perspective on how they can boost the visibility and ranking of their academic information retrieval system profiles. “Greater visibility is implicit in a greater probability of their work being read and cited and, thereby, of boosting authors’ chances to improve their h-index” [8].

By the end of this article [9], it is concluded that *Google Scholar* provides an efficient and scalable

approach to tracking the scholarly output of each research group.

### 3.4.3 *Google Scholar* Evaluation of Journal Articles Impact in Education

The *Using Google Scholar to estimate the impact of journal articles in education* document [6] discusses how *Google Scholar* can be used as a viable alternative to the *WoS* and *Scopus* platforms to evaluate the impact and influence of research output in science education by evaluating the importance of each document Web page through *Google Scholar's PageRank* algorithm.

The *PageRank* algorithm attributes a *PageRank* score to an article. This algorithm relies heavily on, but not entirely, the citations count of each publication [8], “a Web page is considered important if it is linked to by many web pages that are also considered important and if it has few ongoing links to web pages that are not considered important” [6]. The algorithm takes into account both the number of publications that have mentioned a particular work as well as the number of publications that have cited it. Publications that are strongly mentioned by many other publications will have a higher *PageRank* score than publications that are cited by fewer or less significant publications.

The importance of a scientific article is thus assessed by its number of citations and if the articles that have been cited are also classified as important. The articles are later sorted in the research results according to their *PageRank*. The *PageRank* of an article is calculated as the sum of its shares of the *PageRanks* of all the articles that are linked to it. This means that if a document *Y* cites three other documents and one of these documents is document *X*, document *Y* contributes one-third of its *PageRank* score to the *PageRank* score of document *X* [6].

Since the *Google Scholar's* performance evaluations do not involve an excessive number of citations, the *PageRank* algorithm provides an accurate impact assessment of the scientific contribution of each publication.

The *Using Google Scholar to estimate the impact of journal articles in education* document [6] ends its statement by affirming that “*Google Scholar* does a satisfactory job assessing the impact of research output” since it can identify the most influential documents in each sub-field of research. Also, the rate at which *Google Scholar's* citations grew was relatively low in each sub-field of research, meaning that the *Google Scholar* performance evaluations do not involve an excessive number of citations and that the citations from *Google Scholar* provide a reliable measure of impact across sub-fields. Finally, the great majority of citations from *Google Scholar* were from peer-reviewed documents.

## 3.5 *APIs* Description

Our implementation to extract real and updated bibliometric data from authors of *CMU Portugal* is based on harvesting this data from the *Google Scholar's* online pages. We can extract this information through

*Web Scraping.* *Web Scraping* is thus “a technique to extract data from the *World Wide Web (WWW)* and save it to a file system or database for later retrieval or analysis” [16].

Having this in mind, we can use this technique through the use of *APIs*. An *API* is a set of established guidelines that handles data exchanges between systems as an intermediary layer [17]. This allows users to interact with *Hypertext Markup Language (HTML)* pages and to save and use the extracted information. Given this, there are several *APIs* and code libraries that have managed to extract bibliometric data from *Google Scholar*.

### 3.5.1 *SerpAPI*

“*SerpAPI* is a real-time *API* to access *Google* search results” [18]. This tool allows users to extract information from *Google Scholar* without the need of creating a *Web Scraper* from scratch [19]. This *API* works by taking a search query and then returning the data found in a format where the results are separated by title, link, snippet, citations, publication, and other relevant details. Below, we can see an overview of how *SerpAPI* works [18]:

- **API Integration:** Developers integrate the *SerpAPI* into their applications by making *Hypertext Transfer Protocol (HTTP)* requests to the *SerpAPI* server. The *API* provides endpoints for various functionalities, such as searching for specific keywords, retrieving search results, and extracting structured data from *Search Engine Results Pages (SERPs)*.
- **Querying Search Results:** Developers construct queries using the *SerpAPI* parameters to specify the search engine, search query, and additional options like location and language. These parameters help define the context and criteria for the search.
- **Sending Requests:** Developers send a request to the *SerpAPI* server, including the desired parameters in the request *Uniform Resource Locator (URL)* or payload. *SerpAPI* acts as a proxy between the application and the target search engine, making the search request on behalf of the application.
- **Search Engine Parsing:** *SerpAPI* sends the search query to the selected search engine and retrieves the corresponding *SERPs* data. It handles various complexities of search engines, including handling JavaScript rendering, pagination, and different search result formats.
- **Structured Data Extraction:** *SerpAPI* extracts structured data from the search engines results, such as titles, descriptions, *URLs*, featured snippets, images, and other relevant information. This data can be provided in various formats, including *JavaScript Object Notation (JSON)*, *Comma-Separated Values (CSV)*, or *HTML*.



- **Response Handling:** *SerpAPI* collects the extracted data and packages it into an *HTML* response, which is then returned to the developer's application. The response contains structured *SERP* data that can be processed, analyzed, or displayed within the application.

### 3.5.2 *ScraperAPI*

This *API* allows users to scrape data at a great scale [19] without the need to maintain their infrastructure to handle proxies, browsers, and *CAPTCHAs* [20]. This way, users avoid being blocked by domain entities while extracting data from websites. We can see an overview of how *ScraperAPI* works below [20]:

- **API Integration:** To use *ScraperAPI*, developers integrate the API into their applications by making *HTTP* requests to the *ScraperAPI* server. The *API* provides endpoints for various functionalities, such as requesting web page data, managing sessions, and handling proxies.
- **Requesting Web Page Data:** Developers send a request to the *ScraperAPI* server with the *URL* of the web page they want to scrape. *ScraperAPI* acts as a proxy between the application and the target website, making the requests on behalf of the application.
- **JavaScript Rendering:** Many modern websites use JavaScript to dynamically generate content. *ScraperAPI* handles JavaScript rendering by using headless browsers to fully render the page and execute JavaScript code. This ensures that the scraped data includes content loaded by JavaScript.
- **IP Rotation and Proxy Management:** Websites often implement measures to prevent scraping, such as *Internet Protocol (IP)* blocking or rate limiting. *ScraperAPI* rotates *IP* addresses and manages proxies to overcome these challenges. It routes requests through a pool of IP addresses and proxies, preventing *IP* blocks and providing access to websites that would otherwise block scrapers.
- **CAPTCHA Handling:** Some websites may present *CAPTCHA* to verify that the user is not a bot. *ScraperAPI* can automatically handle *CAPTCHAs* by routing the request through an integrated *CAPTCHA*-solving service, allowing the scraping process to continue seamlessly.
- **Response Handling:** *ScraperAPI* retrieves the web page's *HTML* content and returns it to the developer's application in the *HTTP* response. The response includes the scraped data, which can then be processed and utilized within the application.

### 3.5.3 *scrapy* and *scholarly* libraries

*scrapy* [21] and *scholarly* [22] are both open-source and collaborative frameworks that allow users to extract data from websites. These *Python* libraries work by writing code where the user makes a request to *Google Scholar*, where the user specifies which information is to be extracted. Then these libraries return a response with the requested data. The output format of this data is determined by the user. Both libraries send *HTTP* requests to specified *URLs* and retrieves the corresponding responses.

# 4

## Implementation

### Contents

---

4.1 Requirements' Analysis . . . . .	24
4.2 Available Information . . . . .	26
4.3 Tools and Technology . . . . .	30
4.4 System Architecture . . . . .	35
4.5 Methodology . . . . .	36

---

Our implementation is based on extracting data from *Google Scholar* regarding its researchers and publications. This data will be used to quantify the impact caused by the *CMU Portugal* program. The tools and technology used will be explained further on. Our approach was to go through *Ph.D.* and affiliated students of the *CMU Portugal* program. This means that our current solution does not include authors and researchers that are not *Ph.D.* or affiliated students. This way, we can guarantee that the majority of the publications of these students, published during their time at *CMU Portugal*, are within the scope of the *CMU Portugal* program. After we gathered all the needed data, we developed a platform as a dashboard in order to cross-reference and visualize the extracted information.

## 4.1 Requirements' Analysis

In order to evaluate and quantify the impact caused by the *CMU Portugal* program, we will need to analyze the following data:

### 4.1.1 Authors

1. **Affiliations:** The author's affiliation during its participation at *CMU Portugal*. This information is required to identify international collaboration between institutions and countries.
2. **Research Area:** We will need to register each author's research areas in order to identify in which areas is *CMU Portugal* involved.
3. **Citations Count:** The citation count number of each author is the most important value to analyze regarding each researcher. With this value, we are not only able to calculate the number of citations the author has but also to determine the *h-index*, *i10-indexes*, track this value throughout the years, and assess in which years the author peaked and had more influence. These will serve as quantitative metrics regarding the impact caused by the program. Authors with more citations and higher indexes tend to have more effect and impact in their research and field, as well as increasing the *CMU Portugal's* impact over these projects.
4. **Number of Publications:** The number of publications by an author can serve as a quantitative metric to evaluate academic impact because it reflects the productivity and contribution of the author to their field of research. The more publications an author has, the more active they have been in their research and the more they have shared their findings with the academic community.
5. **CMU and PT Advisors:** Each *Ph.D.* or affiliated student has a *CMU* and a *PT* advisor. Having the names of a *PT* advisor (from a Portuguese institution) and a *CMU* advisor (from *Carnegie Mellon University*) can be beneficial when it comes to identifying publications resulting from international collaborations. This is because both advisors represent their respective institutions, and if they are both listed as authors in a publication, this means that there was an international collaboration on that publication.
6. **International Collaborations Count:** As it was previously stated, international collaboration is found by identifying a *CMU* advisor and a *PT* advisor as authors in a publication. Counting how many international collaborations an author has serves as a quantitative metric to calculate the author's international impact.
7. **Student Collaborations Count:** This information is similar to the last metric, but we instead verify if a publication has more than one student as an author. Collaboration is essential to achiev-

ing innovative and impactful results in research, and when students from different universities or countries come together to work on a project, it can lead to an exchange of ideas and methodologies. The number of student collaborations can also serve as a quantitative way to determine the author's impact.

8. **Start and End Research Year:** Each author has a year in which they started their research in *CMU Portugal* and a year in which they stop doing research. We need to determine which author's publications are part of the *CMU Portugal* program and by having this information, we can identify for each author the time period in which all their publications were published during the time they were at *CMU Portugal*. Thus, we can consider that a publication is part of *CMU Portugal* if its publication year is between the author's start of research year and the end of research year, plus one year as a margin.
9. **Other Information:** Besides the information listed above, we also wish to extract the author's name, publications list, *Google Scholar* profile link and profile picture, list of students that have collaborated with the author, and information associated with *CMU Portugal's* current data (type, graduation year, and status).

#### 4.1.2 Publications

1. **Authors:** We require the list of authors in order to identify international collaboration between researchers and identify which students are listed as authors.
2. **Affiliations:** Each author's affiliation will be inherited by the author's publications.
3. **Research Area:** Each author's affiliation will be inherited by the author's publications.
4. **Citations Count:** Once again, the citation count number of each publication is the most important value to analyze and will serve as a quantitative metric regarding the impact caused by the program.
5. **Publication's Type:** Each publication can be either a "*Journal*" ("*article*"), "*Conference Papers*" ("*in proceedings*"), *Ph.D. "Thesis"* and "*Dissertations*", "*Academic Books*" ("*in collection*"), "*Pre-prints*", "*Abstracts*", "*Technical Reports*", and other scholarly literature. Given this, it is important to identify each publication's type. By keeping track of publication types, researchers and evaluators can better understand the research output of individuals and institutions and assess their contributions to their respective fields.
6. **Other Information:** We also wish to extract from each publication the title, year of publication, *Google Scholar* link, *BibTex* and the publication's link and *Digital Document Identifier (DOI)*.

### 4.1.3 Platform

With the information and gathered data listed above, our goal is to implement a platform that makes it possible to create a frontage for a *CMU Portugal*'s "profile" that lists authors/researchers, publications, and citation counts related to the *CMU Portugal*'s scientific research output and tracks its influence. This way, we will be able to visualize and quantify the impact caused by *CMU Portugal* by crossing the data and information of interest above, this will be explained in more detail further. The developed platform was implemented as a visualization dashboard to be used as an internal tool within the *CMU Portugal* team.

## 4.2 Available Information

### 4.2.1 Current *Excel* Data Organization

Currently at *CMU Portugal*, the list of authors with a *Google Scholar* profile page is being saved in a *.xlsx Excel* file, as partially displayed below in *Figure 4.1*. This file is called "*CMU\_PortugalStudents.xlsx*" and currently has 91 entries and contains the following information:

1. **Name:** The author's name
2. ***CMU* and *PT* Advisors:** The names of the student's *CMU* and *PT* advisors with both previous and current advisors.
3. **Start and End Research Year:** Each author's beginning and ending year while doing research for *CMU Portugal*
4. **Graduation Year:** The author's year of graduation.
5. **Status:** The author's status indicates if the author is either a current student, an alumni, or a withdrawn student.
6. **Type:** The type of student indicates if the author is part of the Dual Degree *Ph.D.* program or an affiliated student.
7. **Research Area:** The author's area of research, with both the area's name and acronym.
8. ***Google Scholar* Link:** The author's *Google Scholar* profile page link.

This *Excel* file only contains the profile links to authors that are either *Ph.D.* students or affiliated students. This file served as our starting point to extract data from *Google Scholar*.

	Student Short Name	Start search	Graduation Year	End Search	Status	Google Scholar
1	Alex Gaudio	2018	ongoing	2022	Student	<a href="https://scholar.google.com/citations?user=615F0rKAAAAJ&amp;hl=en&amp;oi=sra">https://scholar.google.com/citations?user=615F0rKAAAAJ&amp;hl=en&amp;oi=sra</a>
2	Alexandre Ligo	2011	2018	2019	Alumni	<a href="https://scholar.google.com/citations?user=zzZu0FAAAAAJ&amp;hl=en">https://scholar.google.com/citations?user=zzZu0FAAAAAJ&amp;hl=en</a>
3	Ana Venâncio	2007	2011	2012	Alumni	<a href="https://scholar.google.com/citations?user=UAJ2g8AAAAJ&amp;hl=pt-PT&amp;oi=sra">https://scholar.google.com/citations?user=UAJ2g8AAAAJ&amp;hl=pt-PT&amp;oi=sra</a>
4	André B. Reis	2010	2017	2018	Alumni	<a href="https://scholar.google.pt/citations?user=7kvefNcAAAAJ&amp;hl=en">https://scholar.google.pt/citations?user=7kvefNcAAAAJ&amp;hl=en</a>
5	André F.T. Martins	2007	2012	2013	Alumni	<a href="https://scholar.google.com/citations?user=mT7ppvwAAAAJ&amp;hl=en">https://scholar.google.com/citations?user=mT7ppvwAAAAJ&amp;hl=en</a>
6	André Regateiro	2008	2014	2015	Alumni	<a href="https://scholar.google.com/citations?user=UC2CrcYAAAAJ&amp;hl=en">https://scholar.google.com/citations?user=UC2CrcYAAAAJ&amp;hl=en</a>
7	Artur Balanuta	2016	ongoing	2022	Student	<a href="https://scholar.google.com/citations?user=LMT0vp4AAAAJ&amp;hl=en&amp;oi=sra">https://scholar.google.com/citations?user=LMT0vp4AAAAJ&amp;hl=en&amp;oi=sra</a>
8	Bernardo Toninho	2009	2015	2016	Alumni	<a href="https://scholar.google.pt/citations?user=LqQoVtgAAAAJ&amp;hl=en">https://scholar.google.pt/citations?user=LqQoVtgAAAAJ&amp;hl=en</a>
9	Brian Swenson	2011	2017	2018	Alumni	<a href="https://scholar.google.com/citations?user=yw55GUAAAAJ&amp;hl=en">https://scholar.google.com/citations?user=yw55GUAAAAJ&amp;hl=en</a>
10	Bruno Vavala	2011	2017	2018	Alumni	<a href="https://scholar.google.com/citations?user=Sz1kTIAAAAAJ&amp;hl=en">https://scholar.google.com/citations?user=Sz1kTIAAAAAJ&amp;hl=en</a>
11	Carla Costa	2007	2013	2014	Alumni	<a href="https://scholar.google.pt/citations?user=kcY6idoAAAAJ&amp;hl=en">https://scholar.google.pt/citations?user=kcY6idoAAAAJ&amp;hl=en</a>
12	Carla Viegas	2016	ongoing	2022	Student	<a href="https://scholar.google.com/citations?hl=de&amp;user=H-lapogAAAAJ">https://scholar.google.com/citations?hl=de&amp;user=H-lapogAAAAJ</a>
13	Chen Wang	2010	2017	2018	Alumni	<a href="https://scholar.google.com/citations?user=JL6WlgAAAAJ&amp;hl=pt-PT&amp;oi=sra">https://scholar.google.com/citations?user=JL6WlgAAAAJ&amp;hl=pt-PT&amp;oi=sra</a>
14	Christian Koehler	2011	2015	2016	Alumni	<a href="https://scholar.google.com/citations?user=ICCMZ8AAAAJ&amp;hl=en">https://scholar.google.com/citations?user=ICCMZ8AAAAJ&amp;hl=en</a>
15	Cláudio Gomes	2021	ongoing	2022	Student	<a href="https://scholar.google.com/citations?user=xlm7eBYAAAAJ&amp;hl=pt-PT">https://scholar.google.com/citations?user=xlm7eBYAAAAJ&amp;hl=pt-PT</a>
16	Cristina Carias	2007	2013	2014	Alumni	<a href="https://scholar.google.com/citations?hl=en&amp;user=M4DYWygAAAAJ&amp;view_op=list_works&amp;sortby=pubdate">https://scholar.google.com/citations?hl=en&amp;user=M4DYWygAAAAJ&amp;view_op=list_works&amp;sortby=pubdate</a>
17	Cristobal Cheyre	2008	2013	2014	Alumni	<a href="https://scholar.google.com/citations?hl=pt-PT&amp;user=3GyrQrMAAAAJ">https://scholar.google.com/citations?hl=pt-PT&amp;user=3GyrQrMAAAAJ</a>
18	Daniel Ramos	2020	ongoing	2022	Student	<a href="https://scholar.google.com/citations?user=AzcMfEAAAAJ&amp;hl=en&amp;oi=sra">https://scholar.google.com/citations?user=AzcMfEAAAAJ&amp;hl=en&amp;oi=sra</a>
19	Diogo Pereira	2021	ongoing	2022	Student	<a href="https://scholar.google.pt/citations?hl=hu&amp;user=7oNscsUAAAAJ">https://scholar.google.pt/citations?hl=hu&amp;user=7oNscsUAAAAJ</a>
20	Dragana Bajovic	2008	2013	2014	Alumni	<a href="https://scholar.google.com/citations?hl=en&amp;user=UwV9710AAAAJ&amp;view_op=list_works&amp;sortby=pubdate">https://scholar.google.com/citations?hl=en&amp;user=UwV9710AAAAJ&amp;view_op=list_works&amp;sortby=pubdate</a>
21	Dušan Jakovetić	2008	2013	2014	Alumni	<a href="https://scholar.google.com/citations?hl=en&amp;user=930h_QAAAAJ&amp;view_op=list_works&amp;sortby=pubdate">https://scholar.google.com/citations?hl=en&amp;user=930h_QAAAAJ&amp;view_op=list_works&amp;sortby=pubdate</a>
22	Eduard Pinconschi	2021	ongoing	2022	Student	<a href="https://scholar.google.pt/citations?hl=en&amp;user=Cs1YV4AAAAJ">https://scholar.google.pt/citations?hl=en&amp;user=Cs1YV4AAAAJ</a>
23	Filipa Reis	2011	2018	2019	Alumni	<a href="https://scholar.google.pt/citations?user=KPi07koAAAAJ&amp;hl=en&amp;oi=sra">https://scholar.google.pt/citations?user=KPi07koAAAAJ&amp;hl=en&amp;oi=sra</a>
24	Filipe Condessa	2011	2016	2017	Alumni	<a href="https://scholar.google.nl/citations?hl=en&amp;user=O6npFoQAAAAJ&amp;view_op=list_works&amp;sortby=pubdate">https://scholar.google.nl/citations?hl=en&amp;user=O6npFoQAAAAJ&amp;view_op=list_works&amp;sortby=pubdate</a>
25	Flávio Cruz	2011	2016	2017	Alumni	<a href="https://scholar.google.nl/citations?user=gh7GUGYAAAAJ&amp;hl=en&amp;oi=sra">https://scholar.google.nl/citations?user=gh7GUGYAAAAJ&amp;hl=en&amp;oi=sra</a>
26	Gabriel Moreira	2021	ongoing	2022	Student	<a href="https://scholar.google.com/citations?hl=en&amp;user=NbrpC8AAAAJ">https://scholar.google.com/citations?hl=en&amp;user=NbrpC8AAAAJ</a>
27	Gopala Anumanchipalli	2008	2013	2014	Alumni	<a href="https://scholar.google.pt/citations?user=VecEj6kAAAAJ&amp;hl=en&amp;oi=sra">https://scholar.google.pt/citations?user=VecEj6kAAAAJ&amp;hl=en&amp;oi=sra</a>
28	Gustavo Gonçalves	2017	ongoing	2022	Student	<a href="https://scholar.google.com/citations?hl=en&amp;user=PdkM60oAAAAJ">https://scholar.google.com/citations?hl=en&amp;user=PdkM60oAAAAJ</a>
29	Hugo Rodrigues	2012	2020	2021	Alumni	<a href="https://scholar.google.com/citations?user=WDug52cAAAAJ&amp;hl=en&amp;oi=sra">https://scholar.google.com/citations?user=WDug52cAAAAJ&amp;hl=en&amp;oi=sra</a>
30	Ivonne Peña	2010	2014	2015	Alumni	<a href="https://scholar.google.com/citations?hl=pt-PT&amp;user=sth5xj8AAAAJ">https://scholar.google.com/citations?hl=pt-PT&amp;user=sth5xj8AAAAJ</a>
31	Jaime Roca	2014	2017	2018	Alumni	<a href="https://scholar.google.nl/citations?user=yIA7NnMAAAAAJ&amp;hl=en&amp;oi=sra">https://scholar.google.nl/citations?user=yIA7NnMAAAAAJ&amp;hl=en&amp;oi=sra</a>


Figure 4.1: *PhD. and affiliate students with a Google Scholar Profile*

## 4.2.2 Google Scholar's Data

Google Scholar is a popular platform that provides access to academic literature, including articles, conference proceedings, books, and more. The platform allows authors to create a profile page that displays their publications, citations, and other academic metrics.

By entering each of the *Google Scholar* profile links on the *Excel* file, we are then able to access each of the authors' profile pages. From each profile, we have access to the author's current affiliation and list of publications. From this list of publications, we can extract both the year when the publication was published and a *hyperlink* that redirects the user to a page with more details about the document (Figure 4.2).

The year of publication is valuable information since it makes it possible to verify if the document is under the time scope in which the author was part of *CMU Portugal*. The *hyperlink* allows us to access additional information about the work, such as citation counts, that can be used to evaluate its impact. From this *link* we can extract the list of authors, the publication's *DOI link* (the *hyperlink connected to the publication's title*), the document's *Portable Document Format (PDF)* file, and the number of citations per year (Figure 4.3).

	<b>Alex Gaudio</b> <a href="#">Carnegie Mellon University</a> Verified email at andrew.cmu.edu - <a href="#">Homepage</a> <a href="#">Explainable Machine Learn...</a> <a href="#">Medical Signal Analysis</a>	<a href="#">FOLLOW</a>
---	---	------------------------

TITLE	CITED BY	YEAR
<a href="#">DeepFixCX: Explainable privacy-preserving image compression for medical image analysis</a> A Gaudio, A Smailagic, C Faloutsos, S Mohan, E Johnson, Y Liu, P Costa, ... Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1495	1	2023
<a href="#">ExplainFix: Explainable spatially fixed deep networks</a> A Gaudio, A Faloutsos, A Smailagic, P Costa, A Campilho Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 13 (2 ...	1	2023
<a href="#">Addressing Chest Radiograph Projection Bias in Deep Classification Models</a> SC Pereira, J Rocha, A Gaudio, A Smailagic, A Campilho, AM Mendonça Medical Imaging with Deep Learning		2023
<a href="#">Explainable Weakly-Supervised Cell Segmentation by Canonical Shape Learning and Transformation</a> P Costa, A Gaudio, A Campilho, JS Cardoso International Conference on Medical Imaging with Deep Learning, 250-260		2022
<a href="#">HeartSpot: Privatized and Explainable Data Compression for Cardiomegaly Detection</a> E Johnson, S Mohan, A Gaudio, A Smailagic, C Faloutsos, A Campilho 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics ...		2022
<a href="#">Privacy-preserving Case-based Explanations: Enabling visual interpretability by protecting privacy</a> H Montenegro, W Silva, A Gaudio, M Fredrikson, A Smailagic, JS Cardoso IEEE Access 10, 28333-28347	6	2022
<a href="#">Explainable Deep Learning for Non-Invasive Detection of Pulmonary Artery Hypertension from Heart Sounds</a> A Gaudio, M Coimbra, A Campilho, A Smailagic, SE Schmidt, F Renna Computing in Cardiology		2022
<a href="#">O-MedAL: Online active deep learning for medical image analysis</a> A Smailagic, P Costa, A Gaudio, K Khandelwal, M Mirshekari, J Fagert, ... Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10 (4 ...	35	2020

**Figure 4.2:** Author's Google Scholar Profile Page

It is also possible to identify, in some publications, what type they are. Directly through *Google Scholar* we can identify three types of publications:

1. **Conference Papers:** This type of publication can also be called “inproceedings”.
2. **Journal:** This type of publication can also be translated as “article”.
3. **Book:** We can also call this type of publication “incollections”



## O-MedAL: Online active deep learning for medical image analysis

[PDF] from arxiv.org

Authors Asim Smailagic, Pedro Costa, Alex Gaudio, Kartik Khandelwal, Mostafa Mirshekari, Jonathon Fagert, Devesh Walawalkar, Susu Xu, Adrian Galdran, Pei Zhang, Aurélio Campilho, Hae Young Noh

Publication date 2020/7

Source Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery

Volume 10

Issue 4

Pages e1353

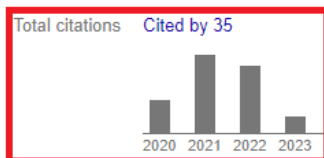
Publisher Wiley Periodicals, Inc

Description Active learning (AL) methods create an optimized labeled training set from unlabeled data. We introduce a novel online active deep learning method for medical image analysis. We extend our MedAL AL framework to present new results in this paper. A novel sampling method queries the unlabeled examples that maximize the average distance to all training set examples. Our online method enhances performance of its underlying baseline deep network. These novelties contribute to significant performance improvements, including improving the model's underlying deep network accuracy by 6.30%, using only 25% of the labeled dataset to achieve baseline accuracy, reducing backpropagated images during training by as much as 67%, and demonstrating robustness to class imbalance in binary and multiclass tasks.

This article is categorized under:

Technologies > Machine Learning

Technologies > Classification ...



Scholar articles [O-MedAL: Online active deep learning for medical image analysis](#)  
A Smailagic, P Costa, A Gaudio, K Khandelwal... - Wiley Interdisciplinary Reviews: Data Mining and ..., 2020  
[Cited by 35](#) [Related articles](#) [All 3 versions](#)

**Figure 4.3:** Publication's Google Scholar Profile Page

We can only identify three types of publications because we extract this information from the publication's profile page, specifically looking for fields labeled as "Conference", "Journal", or "Book". In some cases, the field labeled as "Source" appears instead of the specific publication type. This field does not provide explicit information about the publication type, making it difficult for us to accurately classify these publications.

In the example displayed in *Figure 4.4*, we can identify the highlighted publication's type as a "Conference Paper". In the case of the example in *Figure 4.3*, it was not possible to identify the publication's type since it only displays "Source" in the same position as the publication's type in the examples below. In the case of *Figure 4.3*, we know that the document is an article, and "Source" does not reflect that.

## Explainable Weakly-Supervised Cell Segmentation by Canonical Shape Learning and Transformation

[PDF] from [mlr.press](#)

Authors	Pedro Costa, Alex Gaudio, Aurélio Campilho, Jaime S Cardoso
Publication date	2022/12/4
Conference	International Conference on Medical Imaging with Deep Learning
Pages	250-260
Publisher	PMLR
Description	Microscopy images have been increasingly analyzed quantitatively in biomedical research. Segmenting individual cell nucleus is an important step as many research studies involve counting cell nuclei and analysing their shape. We propose a novel weakly supervised instance segmentation method trained with image segmentation masks only. Our system comprises two models: an implicit shape Multi-Layer Perceptron (MLP) that learns the shape of the nuclei in canonical coordinates; and 2) an encoder that predicts the parameters of the affine transformation to deform the canonical shape into the correct location, scale, and orientation in the image. To further improve the performance of the model, we propose a loss that uses the total number of nuclei in an image as supervision. Our system is explainable, as the implicit shape MLP learns that the canonical shape of the cell nuclei is a circle, and interpretable as the output of the encoder are parameters of affine transformations. We obtain image segmentation performance close to DeepLabV3 and, additionally, obtain an F1-score $F_{IoU=0.5}$ of 68.47% at the instance segmentation task, even though the system was trained with image segmentations.
Scholar articles	<a href="#">Explainable Weakly-Supervised Cell Segmentation by Canonical Shape Learning and Transformation</a> P Costa, A Gaudio, A Campilho, JS Cardoso - International Conference on Medical Imaging with ..., 2022 <a href="#">Related articles</a>  <a href="#">Explainable Weakly-Supervised Cell Nuclei Segmentation by Canonical Shape Learning and Transformation *</a> P Costa, A Gaudio, A Campilho, JS Cardoso - Medical Imaging with Deep Learning, 2022 <a href="#">Related articles</a>

Figure 4.4: Conference Paper's Google Scholar Profile Page

### 4.3 Tools and Technology

The proposed solution takes into consideration the studies conducted by the previous documents [6] [7] [8] [9]. In each of these documents, the *Google Scholar* database was analyzed to see how it could be a measurement tool for research impact and assessment for both institutions and researchers.

In this solution, we extracted the *Google Scholar's* available bibliometric data using the combined use of the *scholarly* [23] and *scrapy* [21] *Python* modules, with the *ScraperAPI* as a proxy service. We used the extracted information to develop a platform that will simplify the online identification process of research published by scientists under the scope of the *CMU Portugal's* international partnership and evaluate the research impact caused by the program.

### 4.3.1 Tools Selection

#### 4.3.1.A Google Scholar

*Google Scholar* was chosen for the following reasons:

- A big majority of the *CMU Portugal*'s publications are conference papers [2], therefore, these have significant relevance when it comes to studying the impact caused by the program through its scientific output.
- The *WoS* and *Scopus* platforms, despite being considered more trustworthy, are more strict with their data and only feature articles and journal publications in their data repositories. This results in these databases having a small number of publications and excluding other forms of research outputs, such as conference papers [6].
- *Google Scholar* features a broader scientific database of research outputs than *WoS* and *Scopus* and does not only consider articles and journals but also gives relevance to conference papers, peer-reviewed documents, theses, books, and book chapters [6].
- The previously referred studies demonstrated that *Google Scholar*'s citations count-based algorithm and researcher's profile are acceptable and accurate tools for assessing an article's and author's scientific contribution and importance [6] [9].
- *Microsoft Academic* was not considered since it was retired on December 31<sup>st</sup> of 2021 according to the *Microsoft Web Page* [24].

#### 4.3.1.B Programming Language Selection

We chose *Python* as a programming language to develop the proposed solution. Python is a high-level programming language that includes a vast standard library and a large number of third-party libraries and frameworks that make it easy to work with different types of data and technologies [25]. A few of the selected tools are *APIs* that come from *Python* libraries, such as *scholarly* [23] and *scrapy* [21]. Given this, *Python* serves as a flexible tool to gather and extract the desired data from *Google Scholar*.

#### 4.3.1.C Proxy APIs Selection

A *proxy* service is a server infrastructure that acts as an intermediary between a client's request for a resource and the server that provides that resource [26]. Since we will be extracting enough data from *Google Scholar* to risk having our *Google Scholar* requests rejected and blocked [22], it's worth emphasizing that we'll need a *proxy* infrastructure. The majority of well-known websites, when they detect that a large amount of information is being extracted from the domain, block the current user's

session, considering that the data is being extracted by *bots* [27]. It is also worth mentioning that free proxy infrastructures will not be considered since they open the possibility for malicious behaviors [28].

Given this, there were considered two *APIs*, *ScraperAPI* and *SerpAPI*. A comparison between the two is displayed in the following table (Table 4.1):

**Table 4.1: *APIs* Tools Selection**

	<b>ScraperAPI</b>	<b>SerpAPI</b>
Open-source	<b>X</b>	<b>X</b>
Price	✓✓	✓✓✓
Simplicity of use	Hard	Mild
Large Scale Data Extraction	✓✓✓	✓
Learning Curve	Hard	Mild
Supported Programming Languages	cURL, Python, NodeJS, PHP, Ruby, Java	cURL, Ruby, Python NodeJS, TypeScript, Go PHP, .NET, Java, Rust, Google Sheets
Documentation	✓✓✓	✓✓

Both *APIs* are not free and have monthly price plans [20], [18]. Considering that we will need to harvest a lot of information from *Google Scholar* we had to consider a trade-off between the monthly price and the amount of data that we can extract with each *API*. Another important factor was how much time it would take to learn how to use each *API*. In the case of *ScraperAPI* we could complement the use of this *Proxy* with the *scrapy Python* library, where we need a “basic understanding of web scraping” [19] and *HTML* elements in order to extract the desired data. This leads to a bigger learning curve when compared to *SerpAPI* is easier to use and gather information.

Despite *SerpAPI* having a smaller learning curve, *ScraperAPI* allows users to extract more data with a cheaper monthly price [19]. Since these *APIs* are rather even in terms of documentation orientation and which programming languages are supported, we reached the conclusion that the *ScraperAPI* proxy service has a better trade-off between the amount of data we can extract and the service’s monthly price. It was shown to be less expensive and allowed us to extract more data than the other considered alternative (*SerpAPI* [18]).

#### 4.3.1.D Platform Tools Selection

**Table 4.2:** Platform Tools Selection

	HTML5/CSS/JS	Outsystems	Microsoft Power BI
Open-source	✓	✓	✓
Simplicity of use	Mild	Easy	Easy
Learning Curve	Mild	Hard	Easy
Data Management Flexibility	✓✓✓	✓	✓✓
Documentation	✓✓✓	✓✓✓	✓✓

In order to develop a dashboard platform to visualize bibliometric data, we evaluated different tools and technologies to choose the best option. Our initial options were *Hypertext Markup Language (HTML5)*, *Cascading Style Sheets (CSS)* and *JavaScript*, *OutSystems*, and *Power Business Intelligence (Power BI)*. We considered the advantages and disadvantages of each tool based on the requirements of the project. We then proceeded to interact with each platform.

*OutSystems* is a low-code platform that offers visual development features and rapid application development. However, we found that *OutSystems* was limited in terms of data manipulation and cross-data analysis. Also, when interacting with this tool, we were concerned about not having much flexibility to program and create the dashboard's desired functionalities [29].

*Power BI*, on the other hand, is a powerful data visualization tool that offers excellent integration with other *Microsoft* products. We found that *Power BI* was easy to use and allowed for the creation of visually appealing dashboards. However, we found that *Power BI* was not as flexible as *HTML5*, *CSS* and *JavaScript* when it comes to customizing and manipulating data [30].

Regarding, *HTML5*, *CSS*, and *JavaScript*, these tools allowed us to have more flexibility in terms of manipulating data and cross-referencing it. The use of these technologies allowed us to customize the dashboard platform to our specific needs. Also, with *JavaScript* we were able to use a library called *D3*.

*Data-Driven Documents (D3)* is an open-source *JavaScript* library used for creating interactive and dynamic data visualizations in web browsers. It provides a wide range of tools for creating custom charts, graphs, and other visualizations. By using this library, we were able to create interactive charts and graphs to better visualize the bibliometric data that we extracted. This allowed us to conduct a more effective analysis and interpretation of the data, as well as provide a more engaging user experience [31].

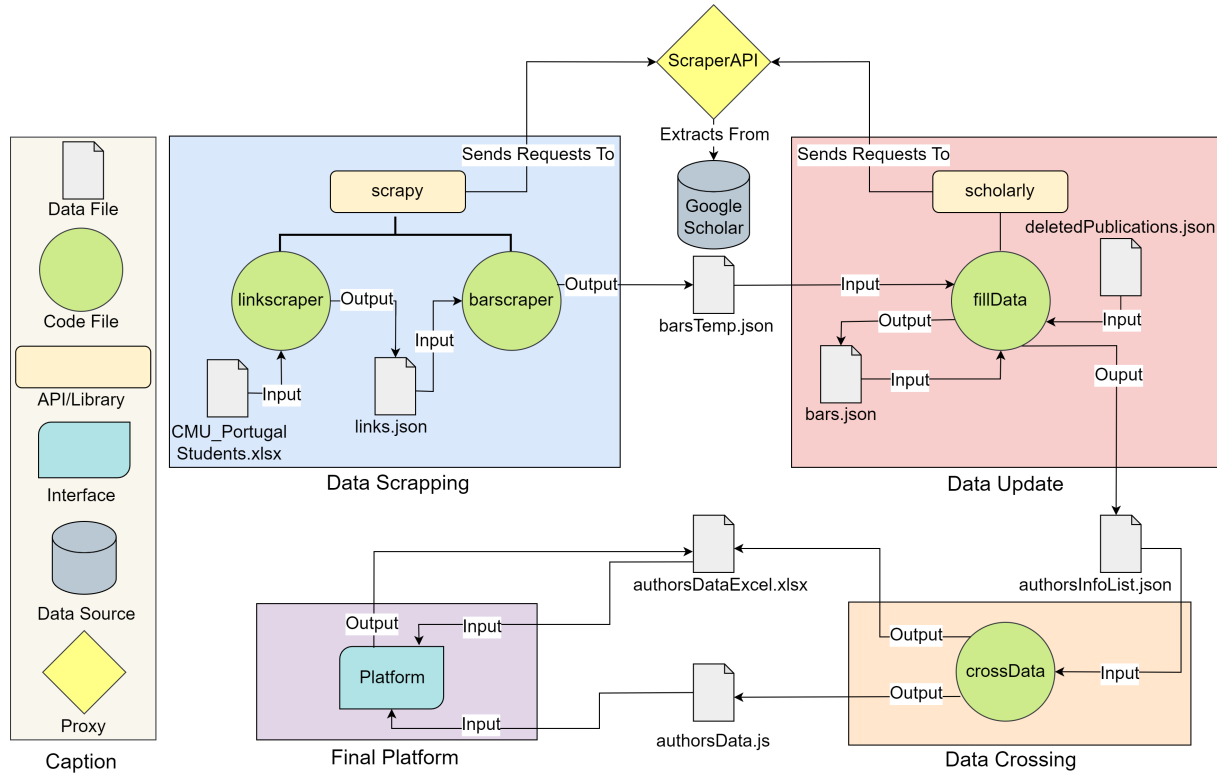
Taking the previous analysis, we chose to use *HTML5*, *CSS*, *JavaScript*, and *D3* since it was the best option for our bibliometric data visualization project due to their flexibility. It allowed us to create a platform that met our specific needs and provided a better user experience compared to the other options we considered. In *Table 4.2* we can see a summarized comparison between each tool.

### 4.3.2 System Dependencies

Given the tools that were chosen and described in the last chapter, our solution has system dependencies of a set of modules and *Python* libraries that needed to be installed:

- **Python:** In this implementation, we used version 3.11.0 [25].
- **Microsoft Visual C++ Build Tools:** A few *Python* libraries have dependencies that can be solved by installing the available build tools from this software development environment [32]. This software is a set of tools and libraries, including a C++ compiler, linker, and libraries, that allows developers to build C++ applications from the command line. This installation is only needed for *Windows* users.
- **lxml:** *lxml* is a *Python* library used for processing *XML* and *HTML* documents. It provides a very efficient and easy-to-use API for parsing and manipulating *XML* and *HTML* data [33].
- **scrapy:** This library is an open-source web scraping framework written in *Python*. It provides a way to efficiently extract structured data from websites and can be used to automate the process of data extraction from web pages. *scrapy* works by sending requests to websites, downloading the *HTML* content, and then parsing the content to extract the data of interest [21].
- **scholarly:** *scholarly* is a *Python* package that provides a simple interface to access and extract publication data from Google Scholar. It allows for easy retrieval of publication information such as titles, authors, abstracts, publication dates, and citation counts. It can also be used to search for publications using keywords and to retrieve a list of publications associated with a particular author. This package uses web scraping techniques to access data from Google Scholar, making it a useful tool for bibliometric analysis [22].
- **xlsxwriter:** This *Python* module is used for writing data to *Excel* files in *.xlsx* format. The module is easy to use and provides a wide range of formatting options [34].

## 4.4 System Architecture



**Figure 4.5:** Information Gathering Architecture

Our system architecture is designed with a modular approach, consisting of four key modules: *Data Scrapping*, *Data Update*, *Data Crossing*, and the *Final Platform*. Each module plays a crucial role in ensuring efficient data extraction from *Google Scholar*, data updating, data processing, and data visualization, respectively. This architecture can be analyzed in *Figure 4.5*.

- **Data Scrapping:** The *Data Scrapping* serves as the foundation, responsible for extracting relevant information from *Google Scholar*. It specifically targets authors listed in the “*CMU.PortugalStudents.xlsx*” *Excel* file, enabling us to gather data that forms the basis of our analysis.
- **Data Update:** The *Data Update* module plays a crucial role in ensuring that our data is up-to-date. It performs several important tasks to keep the information accurate and complete.
- **Data Crossing:** This module receives as input the “*authorsInfoList.json*” file, which contains information about authors and their publications. The data within this file is used to extract valuable insights by crossing and evaluating various data points and fields and to turn this data into information of interest.

- **Final Platform:** The *Final Platform* represents the culmination of our efforts, where we have developed a user-friendly dashboard to visualize the extracted data. By leveraging “*authorsData.js*”, we populate the dashboard with the received data, enabling users to interact with and gain insights from the information in a convenient and intuitive manner. The platform allows for exploring authors’ profiles, publications’ information, citation analysis, and more, providing a view of the academic impact caused by *CMU Portugal*.

## 4.5 Methodology

As referenced in *Section 4.3.1*, this implementation uses the *Python3* programming language [35] with the following libraries and packages:

- **scholarly:** This *Python* package makes it possible to retrieve the author’s and publication’s information from *Google Scholar*. In this library, we can extract the missing publications’ types [23].

**ProxyGenerator:** This package is obtained from the scholarly *API* and it allows scholarly searches to use *proxy* services from *ScraperAPI* [22].

- **scrapy:** This *Python* library is an open-source and collaborative framework that allowed us to extract the needed data from *Google Scholar*. This library allowed us to extract the list of publications of each author and their respective information, such as the number of citations through the years of each publication.

The final platform was implemented by using *HTML5*, *CSS*, *JavaScript* with the *D3* library to generate charts in order to visualize and cross-reference the extracted data from the previous step.

We started our approach by going through the *Google Scholar* links of each author in the *Excel* file provided by *CMU Portugal* (“*CMU\_PortugalStudents.xlsx*”). It is worth noting that this file only includes a list of authors, with a *Google Scholar* profile, that are either *Ph.D.* or affiliated students of the *CMU Portugal* program. This means that our implementation does not include authors outside this scope due to the fact that other researchers currently do not have *Google Scholar* profiles.

### 4.5.1 Data Scraping

This module contains two *Python* code files: *linkspider.py* and *barspider.py*. Each one of these code files is a *spider* created with the *scrapy* library. A *spider* in this context is an agent that can “crawl” along an *HTML* page and extract the necessary information that is chosen by the user [36]. Both these *spiders* use *ScraperAPI* as a *proxy* in each of their requests to *Google Scholar*.



#### 4.5.1.A linkspider

This *spider* is the first step in extracting valuable information from *Google Scholar*. It receives the *Excel* file described in *Section 4.2.1* as input. As previously stated, this *Excel* file contains a list of students who have their own *Google Scholar* profiles. Each entry in the *Excel* file includes a *hyperlink* that directs us to the respective author's *Google Scholar* profile page. This profile contains a list of publications in which the author has participated. The *linkspider* accesses each of the authors' *hyperlinks* and extracts the information of interest from the *Google Scholar* profile (*Figure 4.2*).

In order to identify which publications are under the scope of *CMU Portugal*, they are filtered according to the year when they were published and the author's period window at the *CMU Portugal* program. We only considered publications published under the author's *Start Research* year and the *End Research* year, plus one year as a margin. For instance, if we consider the student "Ana Venâncio" we can observe that she was a student at *CMU Portugal* between 2007 and 2012 (*Figure 4.1*). Given this information, in this example, we only wish to extract information from publications that were published between 2007 and 2013 (2012+1).

From the author's profile page, we extract for each publication the title of the document, the publication year, and the *hyperlink* to the publication's profile page with additional information. Regarding the author, we extract from each profile the author's current affiliation and the author's image. The extracted data is then moved into a *JSON* called "*links.json*". This file contains a list of all publications by every author listed in the "*CMU.PortugalStudents.xlsx*". This means that each entry represents a publication and its respective information. The structure of each publication entry can be visualized in *Figure 4.6*.

```
{
  "author": "Alex Gaudio",
  "affiliation": "Carnegie Mellon University",
  "start_research_year": "2018",
  "end_research_year": "2022",
  "graduation_year": "ongoing",
  "status": "Student",
  "google_scholar_link": "https://scholar.google.com/citations?user=6I5F0rkAAAAJ&hl=en&oi=ao",
  "previous_cmu_advisor": "",
  "cmu_advisor": "Asim Smailagic",
  "previous_pt_advisor": "",
  "pt_advisor": "Aur\u00e9lio Campilho",
  "type": "Dual Degree",
  "research_area": "Electrical and Computer Engineering",
  "research_area_acronym": "ECE",
  "title": "Privacy-preserving Case-based Explanations: Enabling visual interpretability by protecting privacy",
  "year": "2022",
  "link": "https://scholar.google.com/citations?view_op=view_citation&hl=en&user=6I5F0rkAAAAJ&pagesize=100&sortb",
  "image": "https://scholar.googleusercontent.com/citations?view_op=view_photo&user=6I5F0rkAAAAJ&citpid=3",
  "index": 0
},
```

Figure 4.6: "*links.json*" Publication Entry

As we can see in the previous image, the information on each publication is separated by fields:

- **author:** The publication's author name.

- **affiliation:** The publication's author current affiliation.
- **start\_research\_year:** The year when the publication's student started his research at *CMU Portugal*.
- **end\_research\_year:** The year when the publication's author started his research at *CMU Portugal*.
- **graduation\_year:** The year when the publication's student graduated at *CMU Portugal*.
- **status:** Either if a publication's author is a current student at *CMU Portugal* ("*Student*"), a former student ("*Alumni*"), or a student that has not concluded the degree ("*Withdraw*").
- **google\_scholar\_link:** An *hyperlink* to the publication's author *Google Scholar* profile page.
- **previous\_cmu\_advisor:** The names of the publication's author previous *CMU* advisors.
- **cmu\_advisor:** The names of the publication's author current *CMU* advisors.
- **previous\_pt\_advisor:** The names of the publication's author previous *PT* advisors.
- **pt\_advisor:** The names of the publication's author current *PT* advisors.
- **type:** If the publication's author is either an *Dual Degree* student or an *Affiliated* student.
- **research\_area:** The publication's author current field of research.
- **research\_area\_acronym:** The publication's author's current field of research acronym initials.
- **title:** The publication's title.
- **year:** The year when the document was published.
- **link:** An *hyperlink* to the publication's *Google Scholar* profile page.
- **image:** The author's *Google Scholar* profile picture.
- **index:** This *index* indicates the entry position of the publication's author in the *Excel* file. This index is used further in the implementation (Section 4.5.2.D).

If any of the fields are represented as an empty *string* (""), this means that this field is empty. The structure above is repeated in each publication included in the file. The highlighted fields in *Figure 4.6* are extracted from the author's *Google Scholar* profile page (*Figure 4.2*), the remaining fields are from the "*CMU\_PortugalStudents.xlsx*" *Excel* file.

In each of the *CMU* or *PT* advisors' fields (previous and current), if there is more than one advisor, the names are separated by "/" (Example: "advisor1/advisor2"). Also, in this implementation, we considered the research area of each author as the research area of each of his publications. If an author has more than one research area, each research area is also separated by "/" (Example: area1/area2).

The list of publications is “grouped” by their author in alphabetical order. In each author’s “group”, the publications are sorted in descending order according to the year when they were published. This can be observed in *Figure 4.7*.

**Figure 4.7:** “links.json” Publication Order

Since several students can both be authors in the same publication, this file can have duplicate information about the same publication.

#### 4.5.1.B *barspider*

This spider receives as input the “links.json” file generated in the previous spider. For each publication entry, the *barspider* navigates to the publication’s *Google Scholar* page by accessing the *hyperlink* in the “link” field (*Figure 4.6*). From the document’s *Google Scholar* page (*Figure 4.3*), the *barspider* proceeds to extract the document’s title, type, list of authors, the number of citations, and their respective years, the publication’s *link* to the *PDF* file, and the document’s *DOI link*.

After extracting this information, the spider generates as output a file named “barsTemp.json”. Similar to the “links.json” data file, the “barsTemp.json” also includes a list of publications grouped by their author in alphabetical order. In each author group, the publications are once again listed in descending order according to their publication year (*Figure 4.7*).

Once again, each entry represents a publication and its respective data, and the structure of each entry of the “barsTemp.json” can be visualized in the image above (*Figure 4.8*). By analyzing this image, we can observe that the highlighted fields are extracted from the document’s *Google Scholar* page, and the remaining fields are copied from the “links.json” file and represent the same information.

The additional highlighted fields represent the following:

- **authors:** The list of authors that have participated in the document.
- **bars:** The citation values of the publication.
- **years:** The years when each citation’s number value occurred.

```
{
  "author": "Alex Gaudio",
  "affiliation": "Carnegie Mellon University",
  "start_research_year": "2018",
  "end_research_year": "2022",
  "graduation_year": "ongoing",
  "status": "Student",
  "google_scholar_link": "https://scholar.google.com/citations?user=6I5F0rkAAAAJ&hl=en&oi=ao",
  "previous_cmu_advisor": "",
  "cmu_advisor": "Asim Smailagic",
  "previous_pt_advisor": "",
  "pt_advisor": "Aur\u00e9lio Campilho",
  "type": "Dual Degree",
  "research_area": "Electrical and Computer Engineering",
  "research_area_acronym": "ECE",
  "title": "Privacy-preserving Case-based Explanations: Enabling visual interpretability by protecting privacy",
  "publication_year": "2022",
  "authors": "Helena Montenegro, Wilson Silva, Alex Gaudio, Matt Fredrikson, Asim Smailagic, Jaime S Cardoso",
  "bars": [
    "1",
    "2",
    "2"
  ],
  "years": [
    "2021",
    "2022",
    "2023"
  ],
  "pub_type": "article",
  "pub_link": "https://ieeexplore.ieee.org/iel7/6287639/9668973/09729808.pdf",
  "pub_DOI": "https://ieeexplore.ieee.org/abstract/document/9729808/",
  "pub_scholar_link": "https://scholar.google.com/citations?view_op=view_citation&hl=en&user=6I5F0rkAAAAJ&pagesi",
  "image": "https://scholar.googleusercontent.com/citations?view_op=view_photo&user=6I5F0rkAAAAJ&citpid=3",
  "index": 0
},
```

Figure 4.8: “barsTemp.json” Publication Entry

- **pub\_type:** The type of publication (described in Section 4.2.2).
- **pub\_link:** The document’s *link* to the *PDF* file.
- **pub\_DOI:** The *hyperlink* to the publication’s *DOI* page.

Regarding the “bars” and “years” fields, by aligning the elements at the same position, we established a direct relationship between the year and its corresponding number of citations. For instance, in the example on Figure 4.8, we can observe that the field “bars” has “1” in its first position and the field “years” has “2021” in its first position, meaning that in 2021 the publication got a total of one citation.

## 4.5.2 Data Update

The “Data Update” module takes charge of maintaining our data complete, accurate, and up-to-date. It receives the “barsTemp.json” file as input, which serves as the basis for updating the information within it. The update process for this file consists of three distinct stages, each serving a specific purpose:

- **Add Missing Publications:** It identifies any missing publications by comparing the data in “barsTemp.json” with the last updated file named “bars.json”. This comparison allows us to identify old publications that need to be included in our “barsTemp.json”.

- **Filter Deleted Publications:** Checks if any publications listed in *“barsTemp.json”* were previously deleted. This is accomplished by cross-referencing the data with the information stored in *“deletedPublications.json.”* By doing so, we ensure that no deleted publications inadvertently remain in our dataset.
- **Fill Missing Publication Types:** Addresses any missing publication types in *“barsTemp.json.”* It compares the available types in *“bars.json”* and identifies any gaps. To fill in these missing types, the module utilizes the scholarly library, leveraging its resources to obtain the correct publication types for each entry.

The *“bars.json”* is the last updated file, and it serves as a data backup for previously added information. After going through the three stages above and the *“barsTemp.json”* file being updated, this file becomes the *“bars.json”* file, thus becoming the file with the most recent information to date. After this, there is an additional step that reorganizes the structure of the *“bars.json”* file. This step organizes the list of publications by author, allowing for easy access and retrieval of information further on. This organization enhances the overall usability and navigability of our data set. At the end of this step, the information is moved into a file called *“authorsInfoList.json”*.

#### 4.5.2.A Add Missing Publications

Regarding the *Google Scholar* database, the availability of publications is subject to certain changes over time. Occasionally, *Google Scholar* removes publications that it deems irrelevant, while authors themselves may choose to delete their *Google Scholar* accounts or specific publications. An example of this is the student *“Masoud Nazari”* whose *Google Scholar* profile currently no longer exists, but despite this, we were able to extract data from this student before the profile’s deletion. This dynamic poses a challenge when it comes to extracting and preserving data consistently.

To address this challenge and ensure data integrity, a backup mechanism has been implemented. This involves creating a backup file named *“bars.json”* that stores the last updated information obtained from *Google Scholar*. This file has the same structure as *“barsTemp.json”* and it serves as a reference point for future data updates and comparisons.

By preserving the last updated data in *“bars.json”*, we can mitigate the risk of losing valuable information. When we perform subsequent data updates and extract the most recent information into the *“barsTemp.json”* file from *Google Scholar*, we can compare it with the backup file (*“bars.json”*). This comparison enables us to identify any missing publications that might have been deleted or become inaccessible since the last update.

#### 4.5.2.B Filter Deleted Publications

In our implementation, it is possible to delete publications that the user may find irrelevant. This is an external feature that will be explained in more detail in *Section 4.5.5*. Each deleted publication title is added to a file called *“deletedPublications.json”*. When we analyze the *“barsTemp.json”* file, we cross-reference each publication entry with the list of deleted publications in *“deletedPublications.json”*. If there is a match between a publication title in *“barsTemp.json”* and *“deletedPublications.json”*, this means that the publication has been previously deleted. The publication is thus removed from *“barsTemp.json”*.

The *“deletedPublications.json”* file has the following structure:

```
[  
    "Title1",  
    "Title2",  
    "Title3",  
    "Title4",  
    "Title5",  
]
```

**Figure 4.9:** *“deletedPublications.json”* File Structure

Each entry in this file represents the title of a publication that has been deleted.

#### 4.5.2.C Fill Missing Publication Types

As mentioned in *Section 4.2.2*, we are able to identify the document’s type from the majority of publications. This means that in a few publications, it is not possible to extract directly from *Google Scholar* the document’s type. We are only able to identify three types of publications directly from *Google Scholar*.

Since the *scrapy spiders* (*“linkscraper”*) and (*“barscraper”*) extract the data directly from the *Google Scholar* publication’s pages, this means that through this method we are only able to identify if the publication is an *“Journal”* (*“article”*), *“Conference Paper”* (*“inproceedings”*) or a *“Book”* (*“incollections”*). This became an issue since *CMU Portugal* also features publications identified as *“P.h.D. Thesis”* (*“phdthesis”*), *“Tech Report”* (*“techreport”*) or *“Miscellaneous”* (*“misc”*).

To solve this issue, we added to our solution the use of the *scholarly API* [22] with *ScraperAPI* as a *proxy* service. This *Python* library can extract the type of a document from publications that we were unable to do via the *Google Scholar* page.

From some publications whose types were identified by scholarly, we would also filter publications that were identified as *“Master’s Thesis”* (*“masterthesis”*) because these particular documents were likely written during the students’ *Master’s* degree, prior to their involvement with *CMU Portugal*. Since our focus was on publications that were directly associated with *CMU Portugal*, it was important to filter out

these Master's theses to ensure the integrity of our data.

Publications that have no type are represented with "" in the publication type field in *"barsTemp.json"* file. For these publications, we would use *scholarly* to search for its title and return the publication's type. Once this step was complete, we would update the *"barsTemp.json"*. Since this file was now the most updated version, the *"barsTemp.json"* would now become the *"bars.json"* file, thus becoming the most recent backup.

In future updates, to avoid performing more requests through *scholarly* than necessary, before resorting to *scholarly* library to search for missing publication types, we first compare the *"barsTemp.json"* file with the *"bars.json"* file. By comparing these two files, we can identify any publications that have been previously updated by *scholarly*. If a publication has been previously updated, we can confidently conclude that its publication type is accurate and up-to-date, and no further action is needed.

However, for publications that have not been previously updated by *scholarly*, we proceed to use the *scholarly* library to search for the missing publication types.

#### 4.5.2.D Structure Data

```
{
  "author": "Alex Gaudio",
  "affiliation": "Carnegie Mellon University",
  "research_area": "Electrical and Computer Engineering",
  "research_area_acronym": "ECE",
  "type": "Dual Degree",
  "previous_cmu_advisor": "",
  "cmu_advisor": "Asim Smailagic",
  "previous_pt_advisor": "",
  "pt_advisor": "Aur\u00e9lio Campilho",
  "start_research_year": "2018",
  "end_research_year": "2022",
  "graduation_year": "ongoing",
  "status": "Student",
  "google_scholar_link": "https://scholar.google.com/citations?user=6ISF0rkAAAAJ&hl=en&oi=ao",
  "picture": "https://scholar.googleusercontent.com/citations?view_op=view_photo&user=6ISF0rkAAAAJ&citpid=3",
  "index": 0,
  "publications": [
    {
      "title": "DeepFixCX: Explainable privacy\u2010preserving image compression for medical image analysis",
      "pub_year": "2023",
      "author": "Alex Gaudio",
      "authors": "Alex Gaudio, Asim Smailagic, Christos Faloutsos, Shreshtha Mohan, Elvin Johnson, Yuhao Liu, Pedro Costa, Aur\u00e9lio Campilho",
      "bars": [],
      "years": [],
      "pub_type": "article",
      "research_area": "Electrical and Computer Engineering",
      "research_area_acronym": "ECE",
      "pub_link": "https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1495",
      "pub_DOI": "https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1495",
      "pub_scholar_link": "https://scholar.google.com/citations?view_op=view_citation&hl=en&user=6ISF0rkAAAAJ&pagesize=100&sortby=pubdate&citation_for_view=6ISF0rkAAAAJ&citpid=3"
    },
    {
      "title": "Explainable Weakly-Supervised Cell Segmentation by Canonical Shape Learning and Transformation",
      "pub_year": "2022",
      "author": "Alex Gaudio",
      "authors": "Pedro Costa, Alex Gaudio, Aur\u00e9lio Campilho, Jaime S Cardoso",
      "bars": [],
      "years": [],
      "pub_type": "inproceedings",
      "research_area": "Electrical and Computer Engineering",
      "research_area_acronym": "ECE",
      "pub_link": "https://proceedings.mlr.press/v172/costazza/costazza.pdf",
      "pub_DOI": "https://proceedings.mlr.press/v172/costa22a.html",
      "pub_scholar_link": "https://scholar.google.com/citations?view_op=view_citation&hl=en&user=6ISF0rkAAAAJ&pagesize=100&sortby=pubdate&citation_for_view=6ISF0rkAAAAJ&citpid=3"
    }
  ]
}
```

Figure 4.10: *"authorsInfoList.json"* Author Entry

This step is used to reorganize *"bars.json"*. As mentioned, this file is a list of publications grouped by

author in alphabetical order. Within each author's group, the publications are sorted in descending order according to their publication year. This structure remains the same, but the fields are reorganized and a field called "publications" is added per author. This field contains the list of publications for each author. This arrangement made it easier and more efficient to access each of the author's and publication's fields further on in the implementation.

Given this, the information on *"bars.json"* is moved to a file named *"authorsListInfo.json"*. Each entry in this file now represents an author, and each author contains fields about the student and a list of its publications. In each publication, there is also data about the document. The structure of the *"authorsListInfo.json"* file is displayed in *Figure 4.10*.

The successful grouping of publications by the author was made possible by utilizing the *"index"* field in the *"bars.json"* file. This field served as a unique identifier for each author, corresponding to their position in the provided Excel file by *CMU Portugal* (*"CMU.PortugalStudents.xlsx"*). As we scanned through the publications in *"bars.json"*, each publication was already tagged by its author's *index*. Whenever a change in the index number was detected, indicating the transition to another author, we promptly grouped the scanned publications under the current author in the *"authorInfoList.json"* file. This systematic approach allowed us to effectively organize and categorize the publications according to their respective authors.

### 4.5.3 Data Crossing

This module has the objective of generating the final output file that serves as the data source for our platform, namely *"authorsData.js"*. This module takes the *"authorInfoList.json"* file as input, which contains information about the authors and their publications. By leveraging this data, we initiate a series of data cross-referencing operations to extract valuable insights and generate new information for visualization purposes in the final platform.

The *Data Crossing* module is structured into several sequential steps. Firstly, we cross-reference the information about the authors, analyzing their affiliations, research areas, number of publications, and any other relevant attributes. This step allows us to gain a deeper understanding of each author's profile and establish connections between their respective research outputs.

Next, we proceed to cross-reference the publications' data, examining various aspects such as publication types, citation counts, publication years, and any additional relevant bibliometric data. By aggregating and analyzing this information, we gain insights into the authors' research productivity and impact.

Lastly, we employ the collective information from all authors and publications to generate a "profile" that represents *CMU Portugal* as a whole. This profile encapsulates key metrics and characteristics of the research conducted under the scope of the *CMU Portugal* program by crossing the data between



authors and publications, providing a complete view of its scholarly output.

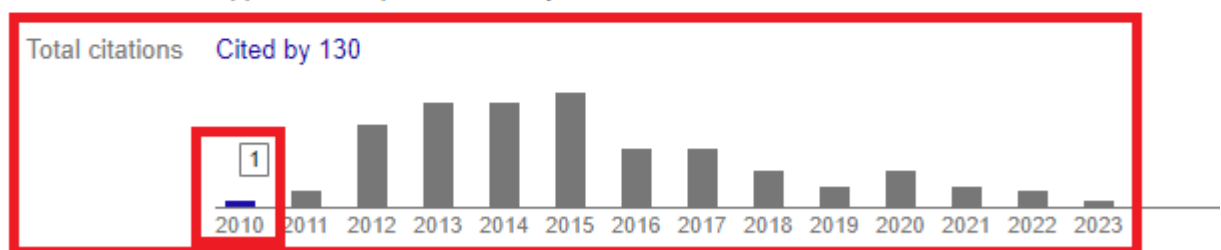
To facilitate data validation and error identification, the information stored in “*authorsData.js*” is also duplicated into an Excel file called “*authorsDataExcel.xlsx*”. This *Excel* format allows for easier data visualization, enabling us to review the data, verify its accuracy, and address any potential errors or inconsistencies before it is integrated into the platform.

#### 4.5.3.A Cross Author’s Data

We initiated the cross-referencing of our data with the authors and their publications.

### An augmented Lagrangian approach to constrained MAP inference.

Authors	André FT Martins, Mário AT Figueiredo, Pedro MQ Aguiar, Noah A Smith, Eric P Xing
Publication date	2011/6/28
Journal	ICML
Volume	2
Pages	2
Description	We propose a new algorithm for approximate MAP inference on factor graphs, by combining augmented Lagrangian optimization with the dual decomposition method. Each slave subproblem is given a quadratic penalty, which pushes toward faster consensus than in previous subgradient approaches. Our algorithm is provably convergent, parallelizable, and suitable for fine decompositions of the graph. We show how it can efficiently handle problems with (possibly global) structural constraints via simple sort operations. Experiments on synthetic and real-world data show that our approach compares favorably with the state-of-the-art.



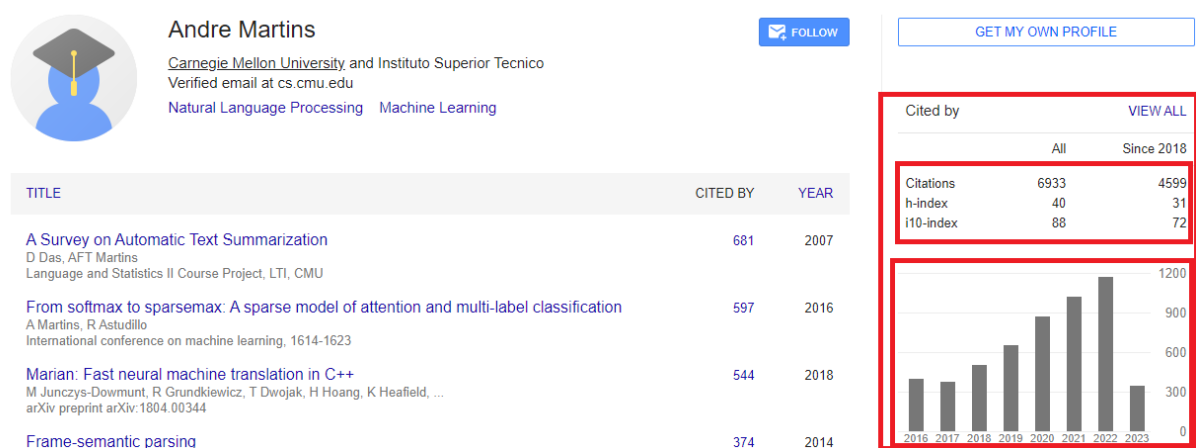
Scholar articles [An augmented Lagrangian approach to constrained MAP inference.](#)  
AFT Martins, MAT Figueiredo, PMQ Aguiar, NA Smith... - ICML, 2011  
[Cited by 130](#) [Related articles](#) [All 23 versions](#)

Figure 4.11: Citations Error Example

**Calculate Metrics through the Number of Citations:** During our data extraction process from *Google Scholar*, we encountered a notable issue pertaining to the recorded citations for certain pub-

lications. It came to our attention that some publications exhibited citations that preceded their actual publication date. Naturally, this was an impossible scenario, as citations cannot occur prior to a publication's existence. One example of this is the publication *"An augmented Lagrangian approach to constrained MAP inference"* written by the student *"André Martins"* can be displayed in *Figure 4.11*. In this example, the publication was published in 2011, but there was a citation in 2010, so this citation was not considered valid information.

To address this anomaly, we implemented a filtering mechanism to discard any citations that occurred before the year of publication. Our objective was to capture only those citations that were logically posterior to the publication date. Consequently, we opted not to extract the displayed number of citations directly from the author's or publication's profiles on Google Scholar. Instead, we focused on extracting and considering the citations per year that were documented after the publication's official publication year.



**Figure 4.12: Citations' Data**

Once this problem was removed, we were able to calculate the same data that is displayed in an author's *Google Scholar* profile (*Figure 4.12*):

- **Total Number of Citations:** By going through all the author's publications and their citations over the years, we calculated the total number of citations by adding each citation number to the respective year of each publication and then repeating this process for all the author's publications while adding the value in each publication. The final calculated value for each author is the sum of all citations in their publications.
- **Total Number of Citations (Last Five Years):** This value was calculated by using the same method described in the previous point, but when we were adding the citations of each year, we only considered the citations of the last 5 years.

- **Number of Citations Per Year:** Once again, we calculate this value with the same method as when calculating the total number of citations, but in the end, instead of adding the publications' number of citations for every year, we saved each citation's number with their respective year. Tracking the number of citations per year provides a dynamic view of an author's research impact over time. It allows us to observe the trajectory of their publications in terms of how widely they are cited by other researchers and scholars. A higher number of citations indicates that the author's work has gained recognition and influence over the years.
- ***h-index*:** The *h-index* represents the number of an author's articles ( $h$ ) that have garnered a minimum of  $h$  citations [37]. To calculate this value, we created a list of citations from the author's publications. We then analyzed the list to identify the minimum number of documents that had received at least  $h$  citations. This minimum value of publications represents the *h-index*, which is a metric used to evaluate an author's impact based on the number of highly cited papers they have. By finding the *h-index*, we gain insight into the author's level of influence and the significance of their research contributions. One example is when we have the following citation values: [5, 4, 2, 1, 1]. In this example, we can determine that the *h-index* is 2 because there are at least 2 documents with at least 2 citations.
- ***h-index (Last 5 Years)*:** To calculate this index, we use the same approach as the last point but only consider publications from the last 5 years.
- ***i10-index*:** This index has the same formula as the *h-index*, but  $h$  has the value 10. This means that we calculate the minimum number of publications with at least 10 citations.
- ***i10-index (Last 5 Years)*:** The same as *i10-index* but only considering publications that were published in the last 5 years.

**Calculate the Number of Publications:** The calculation of the number of publications serves as a valuable metric to quantify an author's productivity. By determining the total number of publications they have produced, we gain insight into the author's research output and the extent of their contributions to academic research. This was calculated from the collected sum of each author's publications.

We can further analyze this data by distributing the number of publications over the years in which they were published. This allows us to examine the author's publication activity and track their research progress over time. By grouping the publications based on their publication year, we can generate a timeline that visualizes the distribution of publications for each author. This timeline provides valuable information about the author's research trajectory, highlighting the years in which they were most active in terms of publishing their work.

Another analysis was conducted to count the number of publications per type and the number of publications per research area for each author. This analysis enables us to identify the predominant

types and research areas of publications and determine the author's primary types of documents and the areas where the author is involved.

**Calculate the Number of International Collaborations:** In our analysis, we thoroughly examine the publication list of each author to identify instances of international collaboration. To determine whether a publication can be considered an international collaboration, we specifically look for the participation of both the author's *CMU* advisor and *PT* advisor as co-authors.

Given this, we calculate the number of international collaborations by adding the sum of the author's publications that are identified as having an international collaboration. By calculating this value, we can indicate that researchers are actively engaging in cross-institutional partnerships, bringing together diverse perspectives, expertise, and resources to advance scientific knowledge.

Also, by grouping each international publication with its publication year, we can perform an analysis of the number of international publications per year for each author. This way, we can study if the number of international collaborations has increased throughout the years since the beginning of the *CMU Portugal* program.

**Calculate the Number of Collaborations Between Students:** In addition to evaluating collaborations between advisors and students, we also considered collaborations between students themselves. By analyzing the authorship of publications, we identified instances where multiple students from *CMU Portugal* were listed as authors of the same publication. This allowed us to quantify the number of collaborations between students.

Calculating the number of collaborations between students is valuable since it provides insights into the level of collaboration and interaction within the *CMU Portugal* community. When students collaborate on research projects or publications, it signifies a willingness to exchange knowledge, ideas, and expertise, which is crucial for fostering a vibrant academic environment.

Additionally, if we group publications that involve collaborations between at least two students according to their publication year, we can analyze the number of student collaboration publications per year for each author. Once again, this analysis enables us to examine whether the frequency of student collaborations has increased over time since the creation of the *CMU Portugal* program.

**Create a *CMU Portugal* "Profile":** With an analysis of the crossed information from the previous points, we can now treat *CMU Portugal* as an authoring entity and calculate various metrics that were previously calculated for individual authors in the same manner. By considering all the authors associated with *CMU Portugal* and their respective publications, we can aggregate the data and evaluate the collective impact of *CMU Portugal* as a whole. We can apply the same metrics used for individual authors, such as the number of publications, citations per year, *h-index*, international collaborations, and other relevant indicators, but now at the program level.

By considering all the publications associated with *CMU Portugal*'s authors as the program's publi-

cations, we can analyze the cumulative impact and contributions made by the collaborative efforts within the program. This approach provides a comprehensive view of the program's research output, collaborative networks, and academic influence. What's more, this created "profile" enables us to assess the program's success in fostering research collaborations and promoting international interactions. Additionally, we also added a list of affiliations for every author. This list allows us to identify which institutions are part of the *CMU Portugal* program. We also counted each author to get a glimpse of how many authors are part of *CMU Portugal*.

#### 4.5.3.B Cross Publication's Data

In the previous points, we performed an individual analysis and crossed data for each author and then for *CMU Portugal* as a whole, creating a "profile" for both the authors and *CMU Portugal*. With the gathered data about the publications, we also created an individual analysis for each publication under the scope of *CMU Portugal*.

Given this, for each publication, we also analyze:

- **Total Number of Citations:** By adding each citation number to the respective year of each publication.
- **Existence of International Collaboration:** By also verifying if there is at least one *CMU* advisor and one *PT* advisor listed as authors in the publication.
- **Existence of Collaboration between Students:** This is verified by the existence of at least more than one student listed as an author in the publication.

**Calculate the number of Students in Student Collaborations:** In addition to analyzing the metrics related to publications, citations, and collaborations of individual authors, we also calculate the number of students who are authors in a publication. This allows us to identify collaboration links between students and gain insights into their research collaborations within the *CMU Portugal* program. The number of students involved in a publication serves as an indicator of collaborative efforts and highlights the extent of interdisciplinary interactions within the program.

#### 4.5.3.C Structure Final Data

After conducting a thorough analysis and cross-referencing all the relevant information as described earlier, we saved the resulting data into a file named "*authorsData.js*," which serves as a data source for our implemented platform. In this case, instead of being a *JSON* file, this file is written in *JavaScript* code, allowing easy access to the stored information. The decision to use this file format was due to the fact that our platform was developed using *HTML5*, and browsers often impose restrictions on foreign

files, such as *JavaScript* or *Python* files, interacting with *HTML* pages due to the *Cross-Origin Resource Sharing (CORS)* policy.

The *CORS* policy is a mechanism that handles cross-origin network access and can restrict the interaction between different file types from separate origins [38]. By converting the data from “*authorsInfoList.js*” into a *JavaScript* file, we were able to circumvent this issue, as *JavaScript* files are not subject to the *CORS* policy limitations. Consequently, the “*authorsData.js*” file seamlessly integrates with our platform, enabling data retrieval and utilization of its data.

```
var CMUData =
{
  "CMUPortugal": {
    "affiliations": [
      "Carnegie Mellon University",
      "Capital One",
      "ISEG- Lisbon School of Economics and Management, Universidade de Lisboa",
      "Ph.D. Electrical and Computer Engineering, Universidade de Aveiro",
      "Carnegie Mellon University and Instituto Superior Tecnico",
      "Tecnico Lisboa and Carnegie Mellon University",
      "Universidade Nova de Lisboa",
      "Pennsylvania State University",
      "Research Scientist, Intel Labs",
      "Maastricht University and IN+, LARSyS",
      "PhD student Carnegie Mellon University",
      "IBM Research",
      "Principal UX Data Scientist, Oracle Inc",
      "Dual-degree Ph.D. Student at Carnegie Mellon University",
      "Merck",
      "PhD Student, Carnegie Mellon University",
      "PhD Student at FCT NOVA",
      "ICONIC center, Faculty of Technical Sciences, University of Novi Sad",
      "Dept. of Math. and Informatics, Faculty of Sciences, Univ. of Novi Sad; COSTNET CA15109",
      "Universidade do Porto",
    ]
  }
}
```

Figure 4.13: “*authorsData.js*” “*CMUPortugal*” Field

Within the *“authorsData.js”* file, we organized the collected information under the *“CMUPortugal”* field (Figure 4.13). This field serves as a container for all the analyzed data pertaining to the *CMU Portugal* “profile”. Additionally, we included a nested field named *“authors”*, which encompasses the comprehensive list of students affiliated with *CMU Portugal* (Figure 4.14). Each individual author within this field contains their respective information and list of publications.

```
"number_of_student_collabs": 17,  
"authors": [  
  {  
    "author": "Alex Gaudio",  
    "affiliation": "Carnegie Mellon University",  
    "research_area": "Electrical and Computer Engineering",  
    "research_area_acronym": "ECE",  
    "collab_students": [],  
    "citations_count": 68,  
    "citations_count5": 68,  
    "citations_per_year": {  
      "2018": 0,  
      "2019": 0,  
      "2020": 8,  
      "2021": 19,  
    }  
  }  
]
```

**Figure 4.14:** *“authorsData.js” “authors” Field*

Within the *“CMUPortugal”* field, we also included a specific field called *“publications.”* This field serves as a list of all the publications associated with *CMU Portugal* (Figure 4.15). The purpose of creating this additional field is to provide convenient access to the information that solely pertains to the publications

themselves, without the need for additional details about the authors at that particular moment. By isolating the publications in this manner, we facilitate the process of retrieving and examining the specific publications.

```
"publications": [
  {
    "title": "DeepFixCX: Explainable privacy preserving image compression for medical image analysis",
    "pub_year": "2023",
    "author": [
      "Alex Gaudio"
    ],
    "authors": [
      "Alex Gaudio",
      "Asim Smailagic",
      "Christos Faloutsos",
      "Shreshta Mohan",
      "Elvin Johnson",
      "Yuhao Liu",
      "Pedro Costa",
      "Aur\u00e9lio Campilho"
    ],
    "citations_count": 0,
    "citations_per_year": {},
    "pub_type": "article",
    "pub_area_tag": "Electrical and Computer Engineering",
    "pub_area_tag_acronym": "ECE",
    "inter_colab": "Yes",
  }
]
```

**Figure 4.15:** *"authorsData.js" "publications" Field*

The data contained in *"authorsData.js"* is also duplicated into an *Excel* file called *"authorsDataExcel.xlsx"* in order to better visualize the data before the final platform was complete. This made it easier to verify if the data within was accurate and correct. A partial view of this file can be observed in *Figure 4.16*. In this image, we can observe information about the authors of *CMU Portugal*. Both *"authorsData.js"* and *"authorsDataExcel.xlsx"* are later passed as input into the final platform.



AUTHOR	AFFILIATION	RESEARCH_AREA	CITATIONS_COUNT	CITATIONS_COUNT_5	H_INDEX	H_INDEX_5
Alex Gaudio	Carnegie Mellon University	Electrical and Computer Engi	66	66	4	4
Alexandre Ligo	Capital One	Engineering and Public Polic	129	110	6	6
Ana Venâncio	ISEG- Lisbon School of Econo	Technological Change and En	4	3	1	1
André B. Reis	Carnegie Mellon University	Electrical and Computer Engi	400	261	9	9
André Cardote	Ph.D. Electrical and Compute	Electrical and Computer Engi	150	45	6	3
André F.T. Martins	Carnegie Mellon University a	Language Technologies	2940	954	26	15
André Regateiro	Carnegie Mellon University	Technological Change and En	117	48	8	4
Artur Balanuta	Técnico Lisboa and Carnegie	Electrical and Computer Engi	220	214	9	9
Bernardo Toninho	Universidade Nova de Lisboa	Computer Science	1045	609	14	14
Brian Swenson	Pennsylvania State Universit	Electrical and Computer Engi	311	241	8	7
Bruno Vavala	Research Scientist, Intel Labs	Computer Science	81	65	4	4
Carla Costa	Maastricht University and IN-	Technological Change and En	21	12	2	2
Carla Viegas	PhD student Carnegie Mellor	Language Technologies	70	70	4	4
Chen Wang	IBM Research	Electrical and Computer Engi	92	66	4	4
Christian Koehler	Principal UX Data Scientist, O	Electrical and Computer Engi	569	399	10	8
Cláudio Gomes	Dual-degree Ph.D. Student at	Electrical and Computer Engi	0	0	0	0
Cristina Carias	Merck	Technological Change and En	227	151	9	8
Cristobal Cheyre	Carnegie Mellon University	Technological Change and En	82	55	5	4
Daniel Ramos	PhD Student, Carnegie Mello	Software Engineering	10	10	2	2
Diogo Pereira	PhD Student at FCT NOVA	Electrical and Computer Engi	8	8	2	2
Dragana Bajovic	ICONIC center, Faculty of Tec	Electrical and Computer Engi	408	131	10	6
Dušan Jakovetić	Dept. of Math. and Informati	Electrical and Computer Engi	1333	737	16	13
Eduard Pinconschi	Universidade do Porto	Electrical and Computer Engi	9	9	1	1
Filipa Reis	Católica Lisbon School of Bus	Technological Change and En	22	12	3	2
Filipe Condessa	Bosch Center for Artificial Int	Electrical and Computer Engi	164	85	6	4
Flávio Cruz	University of Porto and Carne	Computer Science	69	32	3	2
Gabriel Moreira	Carnegie Mellon University	Language Technologies	9	9	2	2
Gopala Anumanchipalli	Assistant Professor, Universi	Language Technologies	199	82	4	4
Gustavo Gonçalves	Carnegie Mellon University a	Language Technologies	9	9	1	1
Gustavo Santos	Department of Physiological	Computer Science	31	9	4	2
Hugo Conceição	CTO at Jumia Services	Electrical and Computer Engi	465	199	12	7
Hugo Rodrigues	INESC-ID, Instituto Superior T	Language Technologies	15	11	2	1
Ivonne Peña	Research Engineer, National	Engineering and Public Polic	39	20	4	3
Jaime Roca	Eindhoven University of Tech	Engineering and Public Polic	388	363	13	13
Jayakorn Vongkulbhisal	Siam Commercial Bank	Electrical and Computer Engi	305	239	8	8
João Diogo de Menezes Falcão	Carnegie Mellon University	Electrical and Computer Engi	313	257	5	5
João G. Martins	PhD student, Computer Scier	Computer Science	450	232	13	9

Global Student Collabs Per Year
Author Data
Author Citations Per Year
Author Publications Per Year
Author Publications Type
Author Publica

Figure 4.16: “authorsDataExcel.xlsx” File

#### 4.5.4 Final Platform

Once the “authorsData.js” file was complete and contained all the relevant information, we proceeded to implement our platform. In accordance with the approach outlined in *Section 4.3.1.A* of our methodology, we used *HTML5*, *CSS*, and *JavaScript*, along with the *D3* library to develop the final platform. We designed our platform to serve as an interactive information dashboard, enabling users to visualize and understand the impact created by *CMU Portugal*.

*HTML5* was used to create the various pages that make up our platform. It provided the necessary structure and layout, allowing us to organize the content and design the user interface. *CSS*, on the other hand, was employed to enhance the visual appearance of each page. By utilizing *CSS* styling rules, we were able to customize the overall aesthetics of the platform. Additionally, we utilized *JavaScript* to write the code that enabled us to manipulate and interact with the data stored in the “authorsData.js” file. With *JavaScript* we were able to dynamically update and display the information, create interactive features, and provide a seamless user experience. To further enrich the data visualization aspect of the platform, we employed the *D3* library. *D3* is a powerful JavaScript library that facilitates the creation of

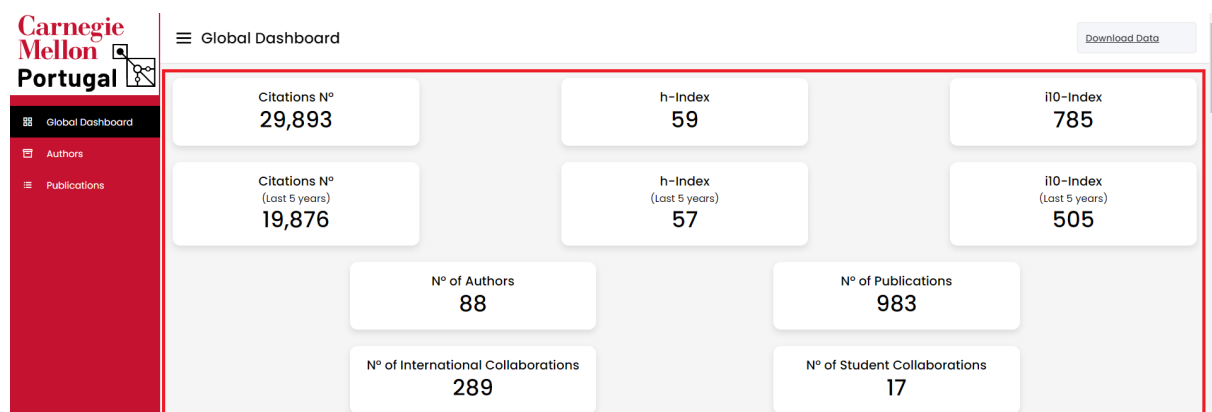
dynamic and interactive data visualizations. With this library, we were able to transform the data from “*authorsData.js*” into visually compelling charts, enabling users to gain meaningful insights and a better understanding of the information presented.

The platform was designed with three functionalities: the “*Global Dashboard*”, “*Authors*”, and “*Publications*”. By separating the platform into these three functionalities, we provide users with a comprehensive and intuitive interface to navigate and explore the data. Whether users are interested in analyzing *CMU Portugal* as a whole, exploring individual authors’ profiles, or delving into specific publications, the platform offers a user-friendly experience. In each of these modules and on every page, it is also always possible to download the visualized content into an *Excel* file. The downloaded file is a copy of the “*authorsDataExcel.xlsx*” file.

#### 4.5.4.A Global Dashboard

The “*Global Dashboard*” serves as the main hub for information regarding *CMU Portugal* as a whole. It provides an overview of key metrics and insights derived from the collective data of all authors and publications associated with *CMU Portugal*. Users can explore the overall productivity and impact metrics of *CMU Portugal* as an academic entity. The “*Global Dashboard*” offers a view of the program’s progress over time, presenting data in the form of charts, tables, and other visual representations.

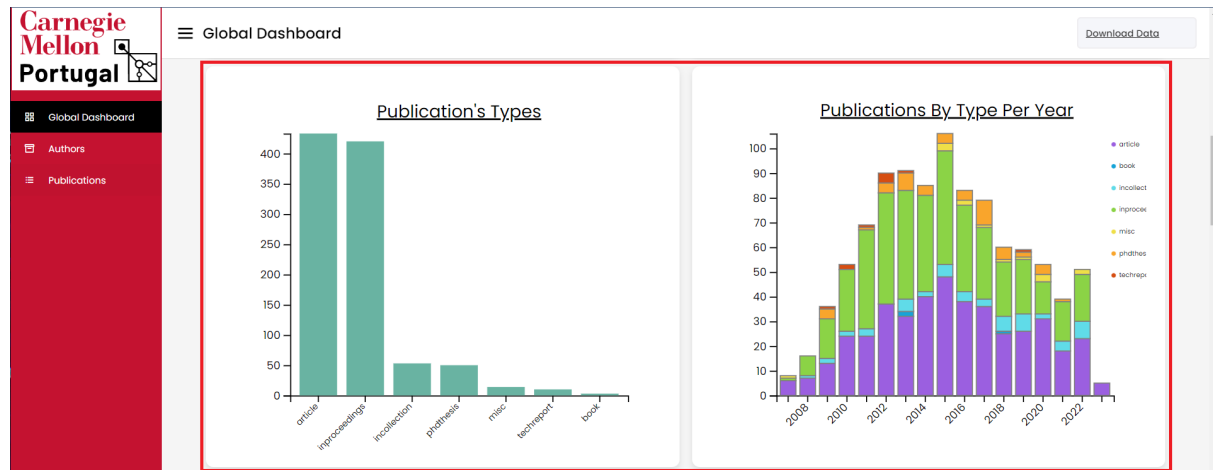
In this functionality, we can visualize quantitative metrics such as the total number of citations, *h-index*, *i10-index*, the total number of authors, the total number of publications, the total number of international collaborations, and the total number of student collaborations. This can be observed in the following image:



**Figure 4.17:** *Global Dashboard's Quantitative Metrics*

The created charts list the following information: citations per year, publications per year, number of publications per type, number of publications by type per year, number of publications per research area, number of publications by research area per year, number of authors per area, publications per

author's research area ratio, international collaborations per year, and student collaborations per year. This information is partially displayed below:



**Figure 4.18:** *Global Dashboard's Charts*

In the case of the charts with the number of publications per type and year and the number of publications by research area and year, there are three variables. In the example of the chart “Publications per Type Per Year” in *Figure 4.18*, each bar represents the number of publications. This bar is separated by color in order to show how many publications have a certain type in their respective years. The chart that analyzes the number of publications by research area and year follows the same logic, but instead of the bars' colors representing the publication types, it represents the publications' research areas.

The chart that evaluates the publications per author's area ratio value is calculated by dividing the number of publications in a certain research area by the number of authors in the same research area. The result is a ratio per research area that gives a more accurate representation of the most common research areas. This chart was necessary due to the fact that, in some cases, there might be more publications in a certain area because an author involved in that area might publish more publications when compared to another author in another area.

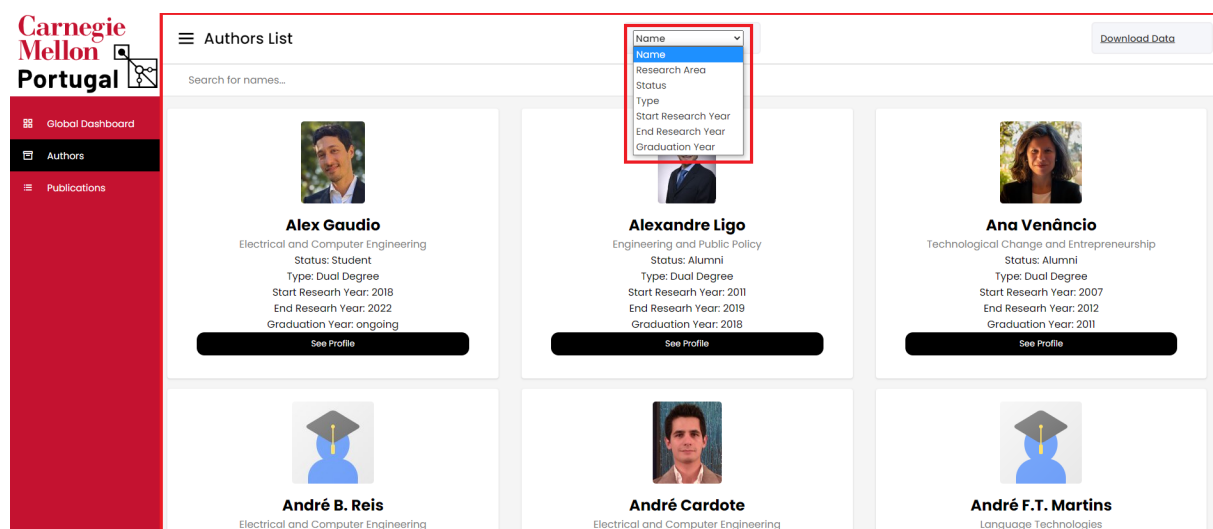
#### 4.5.4.B Authors

The “*Authors*” section of the platform allows users to delve into individual author profiles. Here, users can access detailed information about each author, including their affiliations, list of publications, and academic impact metrics. Users can explore an author's publication history, view collaboration links with other authors, and gain insights into their research productivity and impact. This section offers a granular view of each author's contributions within the *CMU Portugal* program, providing insights into their individual academic journey during their time at *CMU Portugal*.

**Authors List:** In this functionality, the initial page presents a list of authors associated with *CMU Portugal*

(Figure 4.19). This list provides an overview of all the students, allowing users to scroll through the page and locate specific individuals they wish to explore in more detail. Each student is represented by a card containing the author's picture and partial information, giving users a glimpse into their profile.

To enhance the user experience and ease of navigation, the list of authors can be sorted based on various criteria. Users have the flexibility to sort the authors by name, research area, status, type, start and end research years, as well as the year of graduation. This sorting capability enables an efficient organization and exploration of the author list. Additionally, the platform provides search functionality, allowing users to directly search for an author by their name. This feature simplifies the process of finding a specific author, particularly when there are a large number of authors listed.



**Figure 4.19: Author's List Page**

**Author's Profile:** By clicking on the “See Profile” button of a specific student, users gain access to the author's profile (Figure 4.20). This profile is designed to provide a detailed overview of the student's academic journey within the *CMU Portugal* program, presenting information in a similar format to the “Global Dashboard” regarding quantitative metrics and visual charts with individual information regarding the author. In addition to this information, the author's profile includes a dedicated “card” section that showcases personal information relevant to *CMU Portugal*. This includes details such as the student's type, advisors, and research area.

Furthermore, the profile features a table that presents a list of the author's publications (Figure 4.21). Each entry in the table is clickable, leading to the publication's profile page, which provides additional information and insights about that particular publication. The publication table also offers partial information about each publication, and it can be sorted based on various criteria. Users have the flexibility to sort the table by publication year, citation count, publication type, authors, area, international collaboration, and student collaboration.

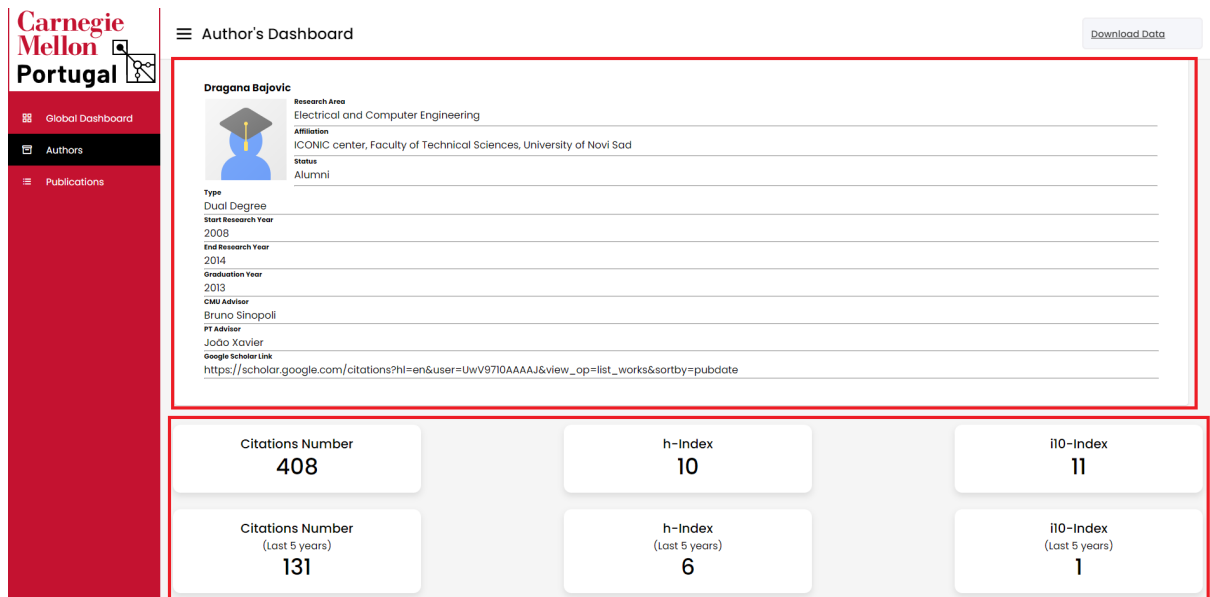


Figure 4.20: Author's Profile Card and Metrics

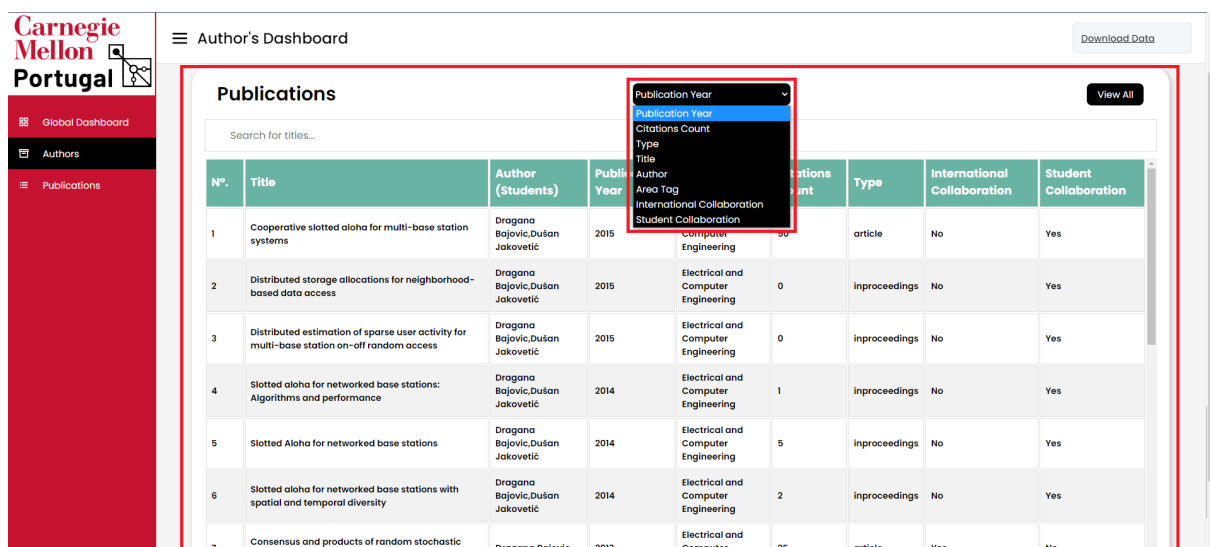


Figure 4.21: Author's Profile Publications List

Additionally, the author's profile includes a table that lists other students who have collaborated with the author (Figure 4.22). It offers sorting mechanisms that allow users to organize the table based on start and end research years, year of graduation, status, type, citation count, number of publications, name, research area, affiliation, number of international collaborations, number of student collaborations, and the number of citations per number of publications ratio. The last filter, the citations per publication ratio, is a calculated value obtained by dividing the total number of citations by the total number of publications. This ratio provides an average measure of the number of citations per publication.

Each entry is clickable, redirecting the page to the author's profile, and displays partial information about each author.

The screenshot shows the 'Author's Dashboard' interface. On the left is a red sidebar with the 'Carnegie Mellon Portugal' logo and navigation links for 'Global Dashboard', 'Authors', and 'Publications'. The main content area is titled 'Author's Dashboard' and includes a 'Download Data' button. Below this is a table of publications. A red box highlights the 'Collaboration Authors' section, which includes a search bar and a table of collaboration data.

Nº	Name	Affiliation	Status	Type	Research Area	Citations Count	Nº of Publications	Start Research Year	End Research Year	Graduation Year	Nº of International Collaborations	Nº of Student Collaborations
1	Dusan Jakovetic	Dept. of Math. and Informatics, Faculty of Sciences, Univ. of Novi Sad; COSTNET CA15109	Alumni	Dual Degree	Electrical and Computer Engineering	1337	27	2008	2014	2013	20	13


**Figure 4.22:** Author's Profile Collaboration Students List

#### 4.5.4.C Publications

The “*Publications*” functionality focuses on providing detailed information about each publication associated with *CMU Portugal*. Users can access details about a publication, including its title, authors, publication type, year, and citation metrics. This section also facilitates filtering and sorting options, allowing users to explore publications based on specific criteria or search for particular titles or authors. The Publications section provides a valuable resource for researchers, enabling them to explore the breadth and depth of academic output within *CMU Portugal*.

**Publications List:** The “Publications” functionality presents users with an initial page featuring a table that includes a list of all publications associated with *CMU Portugal* (Figure 4.23). This table serves as a centralized repository of publications and can be sorted based on various criteria to facilitate easy navigation and analysis. Users have the option to sort the table by publication year, citation count, publication type, title, authors, research area, international collaboration, student collaboration, and contributing advisors. The table provides partial information for each publication, offering a snapshot of key details.

**Publication's Profile:** By clicking on an entry in the table, users can access the publication's profile, which provides a more in-depth view of the publication. Similar to the authors' profiles, the publication's profile includes an information “card” that contains specific details about the publication sourced from *Google Scholar*.



Publications List

Publication Year


Download Data

Search for titles...

Nº.	Title	Author	Publication Year	Area Tag	Authors	Citations Count	Type	International Collaboration	CMU Advisor	PT Advisor
1	DeepFixCX: Explainable privacy-preserving image compression for medical image analysis	Alex Gaudio	2023	Electrical and Computer Engineering	Alex Gaudio, Asim Smalagic, Christos Faloutsos, Shreshtha Mohan, Elvin Johnson, Yuhao Liu, Pedro Costa, Aurélio Campilho	0	article	Yes	Asim Smalagic	Aurélio Campilho
2	Evaluating gesture-generation in a large-scale open challenge: The GENEA Challenge 2022	Carla Viegas	2023	Language Technologies	Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, Gustav Eje Henter	0	article	No		
3	Fill in the blank for fashion complementary outfit product Retrieval: VISUM summer school competition	Gabriel Moreira	2023	Language Technologies	Eduardo Castro, Pedro M Ferreira, Ana Rebelo, Isabel Rio-Torto, Leonardo Capozzi, Mafalda Falcão Ferreira, Tiago Gonçalves, Tomás Albuquerque, Wilson Silva, Carolina Afonso, Ricardo Gamelas Sousa, Claudio Cimorelli, Nadia Daoudi, Gabriel Moreira, Hsiu-yu Yang, Ingrid Hrga, Javed Ahmad, Monish Keswani, Sofia Beco Yangyi Zhao, Yunsik	0	article	No		

Figure 4.23: Publications List Table

In addition to the information “card”, the publication’s profile also presents quantitative information, like the total number of citations, and a chart depicting the number of citations over the years provides a visual representation of the publication’s citation trajectory, enabling users to assess its long-term impact. The publication’s profile further includes a table listing the students who contributed to the publication. This table offers valuable information about the students involved in the publication. A partial view of this profile page can be observed in (Figure 4.24).



Publications's Dashboard

Download Data

Title

Matrix completion for weakly-supervised multi-label image classification

Year

2014

Full Authors List

Ricardo Cabral, Fernando De la Torre, Joao Paulo Costeira, Alexandre Bernardino

Type

article

Area Tag

Electrical and Computer Engineering

International Collaboration

Yes

CMU Advisor

Fernando De la Torre

PT Advisor

Joao Paulo Costeira

Student Collaboration

No

Link

[http://www.isr.tecnico.ulisboa.pt/larsys2017/Cabral\\_et\\_al.pdf](http://www.isr.tecnico.ulisboa.pt/larsys2017/Cabral_et_al.pdf)

DOI Link

<https://ieeexplore.ieee.org/abstract/document/6866218/>

Google Scholar Link

[https://scholar.google.com/citations?view\\_op=view\\_citation&hl=pt-PT&user=NB9xXQcAAAAJ&pagesize=100&sortby=pubdate&citation\\_for\\_view=NB9xXQcAAAAJjCSPb-OGe4C&hl=en](https://scholar.google.com/citations?view_op=view_citation&hl=pt-PT&user=NB9xXQcAAAAJ&pagesize=100&sortby=pubdate&citation_for_view=NB9xXQcAAAAJjCSPb-OGe4C&hl=en)

Delete Publication

Citations Number

196

Figure 4.24: Publications' Profile Page

### 4.5.5 External Features

In addition to the platform itself, we have developed two external features that enhance the functionality of our system. These features are not directly integrated into the platform but provide valuable capabilities for users.

The first feature allows users to run the code for the “*Data Extraction*”, “*Data Update*”, and “*Data Crossing*” modules by simply clicking on an executable file. This file opens an interface where users can select which part of the code they wish to execute. The executable file simplifies the process and provides a user-friendly interface for executing the desired code.

The second feature, as briefly mentioned in *Section 4.5.2*, allows users to eliminate publications by indicating one or more publication titles. This functionality is also implemented using an executable file, which opens an interface where users can input the titles of the publications they wish to exclude. The executable file streamlines the process of excluding publications and makes it more accessible to users.

Both features utilize executable (.exe) files, which are designed to run on *Windows* operating systems. For other operating systems, users can execute the interface’s *Python* code by entering a command line in a terminal console. The interfaces for both features were developed using the “*tkinter*” library, which is a popular *Python* library for creating graphical user interfaces [39]. By leveraging “*tkinter*”, we were able to design intuitive and user-friendly interfaces for these external features.

Initially, we intended to integrate these features directly into the platform. However, due to the *CORS* policy limitations, we were not able to use external code files to interact directly with the dashboard [38]. For this reason, we opted to develop both features as separate executable files that could be easily executed with the assistance of an executable file and graphic interfaces. This approach allows users to run the features independently without the need to open a terminal console and execute *Python* commands.

#### 4.5.5.A Run the Modules Code Feature

Our implementation follows a modular approach, as discussed in *Section 4.4*, where, while excluding the “*Final Platform*” module, we divided the process into three main roles: data extraction, data update, and data crossing. To provide users with more flexibility and control over the execution of these modules, we developed a feature that opens an interface allowing them to select the specific modules they wish to execute.

The modular approach enables users to selectively execute different parts of the code based on their needs. Not every user may require the complete functionality at once, and by offering the option to choose specific modules, they can customize the execution to suit their requirements. For example, users can choose to execute the data extraction module alone to retrieve new data without performing the data crossing or update steps. Additionally, this feature accommodates scenarios where some files



have already been updated and users only need to write the final data to the platform. This saves time and computational resources, especially when dealing with large data sets or when only specific data components need to be refreshed.

We can observe this feature’s graphical interface in *Figure 4.25*:

**Figure 4.25:** *Modules' Code Graphical Interface*

In *Figure 4.25* we can observe that the user can select either “Yes” or “No”, indicating which modules the user wished to run.

#### 4.5.5.B Delete Publications

**Figure 4.26:** *Delete Publications Graphical Interface*

We developed the “*Delete Publications*” feature to address a potential issue that arises due to the way we consider publications within the scope of *CMU Portugal*. When determining which publications to include, we consider publications between the author’s start research year and the end research year, with an additional one-year margin. This approach accounts for the possibility that some publications

may have been delayed in their publication timeline.

However, this approach also introduces the possibility of including publications that are no longer relevant to *CMU Portugal*. Once an author's research period ends, they are no longer officially affiliated with *CMU Portugal*. Consequently, any publications they produce beyond the end research year may not fall within the program's scope.

To ensure data accuracy and maintain the integrity of the *CMU Portugal* profile, we developed the "Delete Publications" feature. This feature allows users to indicate specific publications that should be excluded from the *CMU Portugal* dataset. By providing this option, we enable users to remove publications that fall outside the intended timeframe, eliminating any potential inaccuracies or misrepresentations. The graphical interface can be visualized in *Figure 4.26*:

This feature receives both as input and output the *"deletedPublications.json"* file. Users can simply copy the desired publication titles and paste them into the designated text box (highlighted in *Figure 4.26*) within the interface. Each publication title should be separated by a comma (",") to ensure proper formatting. Once the user has inserted the publication titles, the underlying code takes over. It adds individual titles to the *"deletedPublications.json"* file. This file serves as a record of the publications that have been marked for deletion from the *CMU Portugal* dataset. After this, it executes the code of the *"Data Update"* and *"Data Cross"* modules, putting the data up-to-date.

# 5

## Evaluation

### Contents

5.1 User Characterization . . . . .	64
5.2 Platform Validation . . . . .	65
5.3 Tasks Execution Results . . . . .	67
5.4 Usability Questionnaire Results . . . . .	68
5.5 Results Discussion . . . . .	70

For our evaluation process, we adopted the evaluation method outlined in the “*The Development of Heuristics for Evaluation of Dashboard Visualizations*” article [40]. In this document, the authors developed a “heuristic evaluation checklist that can be used to evaluate systems that produce information visualizations” [40]. It combines the principles of Nielsen’s heuristics with heuristic principles developed by previous researchers specifically designed to evaluate information visualization.

The evaluation process consisted of three stages. First, we created a questionnaire to collect demographic data from the users participating in the evaluation. This questionnaire helped us gather user information, including their background and experience with similar platforms. Next, the users were guided through a series of pre-defined tasks that aimed to familiarize them with the platform’s functionalities and features. These tasks were carefully designed to ensure that users explored all aspects of

the platform, enabling us to assess its usability.

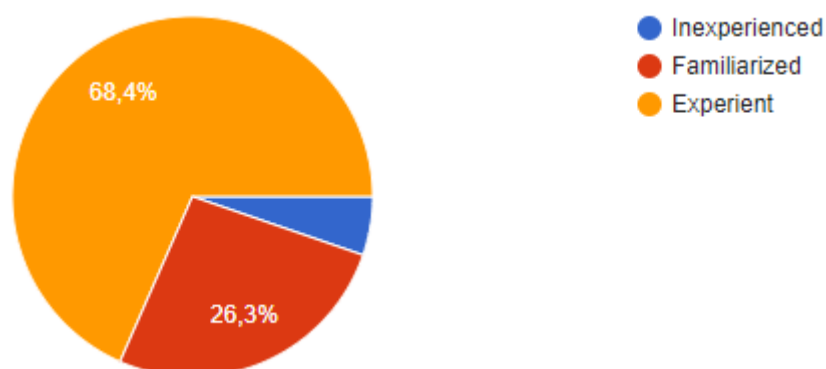
Finally, we asked users to answer a usability questionnaire incorporating the heuristic checklist from “The Development of Heuristics for Evaluation of Dashboard Visualizations” article. By employing this checklist, we were able to evaluate the user’s satisfaction level with the platform and identify any areas that required improvement. In this evaluation, we excluded the assessment of the external features (*Figure 4.5.5*) as they do not directly contribute to the usability of the platform. External features refer to specific functionalities or tools that are designed for the requirements of *CMU Portugal* stakeholders.

## 5.1 User Characterization

During this stage, we conducted structured interviews with 19 participants. We provided each participant with a questionnaire to collect demographic data (*Appendix A*). Our target users for the evaluation of the platform were individuals who had previous experience searching for academic literature and using search engines for academic content, such as *Google Scholar*.

Among the participants, 47% were currently studying in academic programs, with ages ranging from 22 to 30 years old. Out of this group, 53% identified as female, while 47% identified as male. In terms of educational background, 68% had completed a bachelor’s degree, and 32% had completed a master’s degree.

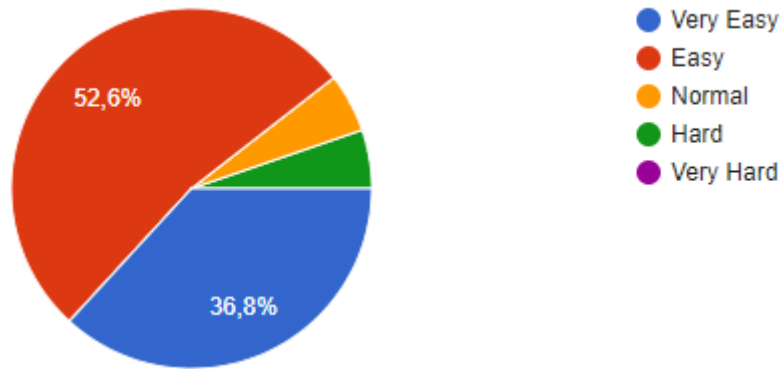
When assessing their experience with technological devices, 68% of the participants considered themselves to be experienced and 26% considered themselves to be familiar (*Figure 5.1*).



**Figure 5.1:** User's Experience Circular Chart

Additionally, when asked about their ease in learning new technological devices or applications, 37% responded that it was very easy, while 53% stated that it was easy (*Figure 5.2*).

It is important to note that all participants in this group had previous experience searching for academic literature and utilizing search engines for academic content. Also, 74% of the participants are currently or were enrolled in “*Computer Science and Engineering*” and had previous experience with



**Figure 5.2:** *User's Easiness Circular Chart*

user usability and interface development. This familiarity with academic literature search processes provided them with a foundation for evaluating the platform's features and capabilities.

## 5.2 Platform Validation

During the evaluation process, we provided each participant with a user guide (*Appendix B*) that outlined three specific tasks to be performed on the final platform. These tasks were designed to cover the main functionalities of the platform, the “*Global Dashboard*”, “*Authors*”, and “*Publications*”. The user guide included in each task a step-by-step list of instructions to ensure that participants fully explored each functionality. Each task was timed, and our objective was to ensure that, on average, users could complete all three tasks within a time frame of 10 minutes [41]. This was to simulate a long visit to a website or platform where users may have limited knowledge about the interface and functionalities. This way, we are able to evaluate the intuitiveness and ease of use of our platform.

Once the participants completed the tasks outlined in the user guide, we asked them to answer a usability questionnaire (*Appendix C*). The questionnaire was based on a set of heuristics derived from the article “*The Development of Heuristics for Evaluation of Dashboard Visualizations*” [40]. For each heuristic, participants were asked to rate their agreement on a scale of 1 to 5, with 1 indicating total disagreement and 5 indicating total agreement, regarding how well the dashboard adhered to each heuristic's description. The list of heuristics is presented below:

1. **Visibility of System Status:** The system should always update the user on what is happening by providing suitable feedback in a timely manner.
2. **Match between System and the Real World:** Instead of using system-oriented jargon, the system should employ words, phrases, and concepts that are known to the user. Adhere to etiquette, arranging facts in a logical and natural sequence.

3. **User Control and Freedom:** Instead of letting the system choose and order tasks for the user (when appropriate), this should be left up to the user. Users will require an unmistakably designated “emergency exit” in order to quit the undesirable state without having to engage in a lengthy discourse. Users should decide for themselves how much it will cost to stop doing something.
4. **Consistency and Standards:** Users shouldn’t have to question whether various terms, circumstances, or behaviors mean the same thing.
5. **Recognition rather than Recall:** Make options, actions, and objects obvious. It shouldn’t be necessary for the user to remember details from one section of the dialogue to the next. When necessary, instructions for using the system should be readily available or apparent.
6. **Flexibility and Efficiency of Use:** When it comes to choosing how to discover content, the system should give users many possibilities. Users should be able to effectively accomplish their objectives.
7. **Aesthetic and Minimalist Design:** Information that is unnecessary or rarely used shouldn’t be included in dialogues. Each additional piece of information in a conversation competes with the pertinent pieces and reduces their relative exposure.
8. **Spatial Organization:** Relates to how a visual representation is organized overall, how simple it is to find specific information items in displays, and how elements are distributed in representations.
9. **Information Coding:** The use of symbols or representations to facilitate perception
10. **Orientation:** Providing assistance to the user and guiding them in the visualization

After participants completed the rating of each heuristic, we sought to gather more specific feedback by asking if they encountered any usability issues or problems with the dashboard. If the response was affirmative, participants were then prompted to rate the severity of the identified issue on a scale of 1 to 4. This severity scale is explained in more detail below [42]:

- **1 - Aesthetic Problem:** Does not need to be corrected.
- **2 - Minor Usability Issue:** Can be corrected, but it is not urgent to do so.
- **3 - Major Usability Issue:** It is important to correct the issue.
- **4 - Usability Catastrophe:** It is imperative to correct the issue.

## 5.3 Tasks Execution Results

As previously mentioned, we recorded the time taken by each user to complete each task. By capturing this data, we were able to analyze and evaluate the efficiency and effectiveness of the platform's usability. The recorded times allowed us to calculate the average time spent by users on each task and overall.

In the *Table 5.1* below, we can observe the recorded times for each task and user:

**Table 5.1:** Recorded Tasks' Times

User	Measured Time (MM:SS)			
	Task 1 (Global Dashboard)	Task 2 (Authors)	Task 3 (Publications)	Total
1	02:58	02:29	01:48	07:05
2	03:06	03:11	01:15	07:32
3	02:50	04:48	02:05	09:43
4	04:44	05:20	02:39	13:43
5	04:05	02:58	01:11	08:14
6	08:56	06:30	02:59	13:43
7	02:20	05:20	01:07	08:47
8	08:00	05:23	03:14	16:37
9	03:46	05:27	04:13	13:26
10	03:20	02:14	01:14	06:48
11	05:17	06:51	03:11	15:19
12	04:11	05:11	03:11	12:33
13	02:45	02:20	01:19	06:24
14	04:02	04:50	01:25	10:17
15	03:37	02:16	02:00	08:53
16	02:54	02:12	01:16	06:22
17	01:22	01:40	01:25	04:27
18	02:11	02:48	01:22	06:21
19	07:14	07:06	01:53	16:13
<b>Average</b>	04:05	04:09	02:02	10:08

## 5.4 Usability Questionnaire Results

**Table 5.2:** Heuristics Results

	Heuristics									
User	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
1	4	4	4	4	4	4	4	4	4	4
2	5	5	4	5	5	5	4	3	5	5
3	5	5	5	5	5	5	5	5	4	5
4	5	5	5	5	5	5	5	5	5	5
5	5	5	3	5	5	5	5	4	5	4
6	4	5	5	5	4	4	5	5	5	4
7	4	5	4	4	5	5	5	5	4	4
8	3	3	4	4	2	2	2	1	2	2
9	5	5	5	5	5	5	5	5	5	5
10	5	5	5	5	3	4	5	5	4	5
11	5	5	5	5	5	5	5	5	5	3
12	4	5	2	5	5	5	5	5	5	4
13	5	5	4	5	5	5	5	5	4	4
14	4	5	4	4	5	5	5	5	5	4
15	4	3	3	5	3	4	4	4	3	2
16	3	4	5	5	2	5	4	4	5	3
17	4	5	4	4	5	5	5	5	5	4
18	4	5	4	5	5	5	5	4	5	4
19	4	5	1	3	3	4	2	2	3	4
<b>Average</b>	4.3	4.7	4	4.6	4.3	4.6	4.5	4.3	4.4	3.95
<b>Results (%)</b>	86	94	80	92	86	92	90	86	88	79

During the evaluation process, we asked users to rate their agreement with each heuristic on a scale of 1 to 5, where 1 represented “Totally Disagree” and 5 represented “Totally Agree.” This rating system allowed us to assess how well the platform respected each heuristic.

To analyze the results, we calculated the average rating for each heuristic across all users. We also computed the average rating percentage for each heuristic by dividing the average rating by the maximum score of 5. This percentage represents how well the platform adhered to each heuristic, with higher percentages indicating better alignment.

These results can be observed in *Table 5.2*. Each heuristic column is represented as an acronym

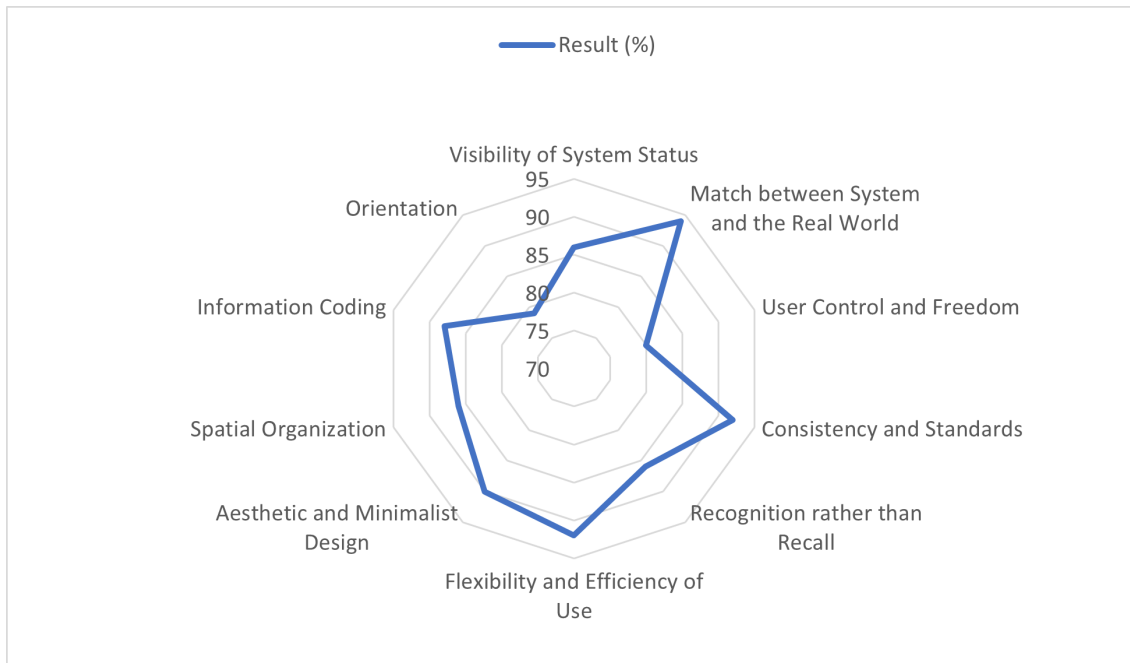


in the following manner: “*Visibility of System Status (H1)*”, “*Match between System and the Real World (H2)*”, “*User Control and Freedom (H3)*”, “*Consistency and Standards (H4)*”, “*Recognition rather than Recall (H5)*”, “*Flexibility and Efficiency of Use (H6)*”, “*Aesthetic and Minimalist Design (H7)*”, “*Spatial Organization (H8)*”, “*Information Coding (H9)*”, “*Orientation (H10)*”.

Regarding usability issues, 74% of users have found problems with our dashboard and suggested the following improvements:

- **Back Button:** Currently, the platform uses the browser's back button to return to previous pages. Six participants have suggested adding a back button for higher flexibility. On average, this issue received a rating of 2.
- **Breadcrumbs:** Two users specifically suggested the implementation of breadcrumbs. Breadcrumbs provide a visual representation of the user's location within the platform's hierarchy, allowing them to see the path they have taken to arrive at their current page. On average, this issue received a rating of 2.
- **Lateral Index Navigation:** We received valuable feedback from participants regarding the need for improved navigation. Specifically, eight participants suggested the addition of a navigation index in the lateral navigation bar. This issue arose when certain pages within the platform displayed content that exceeded the visible area, resulting in users not being able to fully explore all the features available. To address this concern, participants recommended the inclusion of a navigation index that would indicate the complete information available on each page. On average, this issue was rated with a severity classification of 2 (2.1).
- **Clickable Elements:** Participants have highlighted this issue because, when interacting with tables, users had difficulty noticing that each entry was a clickable element. This would happen with other clickable elements that were not highlighted as such. For this reason, participants have suggested adding visual clues that would indicate that an element is clickable. This issue received a rating of 2 based on 5 answers.
- **Category Chart Colors:** One of the interviewed users noticed that in a few charts where each bar would represent a kind of category (like the publication's type) that every bar displayed the same color. The suggested improvement was to add different colors to each bar in charts with categorical representation. This issue received a classification of 2.
- **Table Filters:** Currently on our platform, we can only sort the displayed tables by one filter at a time. Three participants suggested that the tables could support more than one filter at a time to facilitate navigation through the table. This problem received an average score of 3 (2.7).

## 5.5 Results Discussion



**Figure 5.3:** *Heuristic Evaluation Results*

Based on the evaluation results, we can conclude that the execution of tasks on our platform was generally acceptable. The average time taken by users (5.1) to explore the entire dashboard was 10 minutes and 8 seconds. Although this is slightly above the target of 10 minutes, we can conclude that users were able to navigate and complete the tasks within a reasonable timeframe. This is because, during the tasks' execution, users had to read and understand the user guide while they performed each task, adding a small fraction of time to the overall duration. This suggests that our platform has a user-friendly and intuitive design. Users were able to grasp the layout and functionality of the platform, enabling them to accomplish their tasks.

Regarding the usability questionnaire results and *Figure 5.3*, our main issues were related to “*Orientation*”, “*User Control and Freedom*”, “*Flexibility and Efficiency of Use*”, “*Visibility of System Status*”, “*Information Coding*” and “*Spatial Organization*”. In terms of “*Orientation*”, users have identified the issue of “Breadcrumbs”, and “Lateral Index Navigation”. When it comes to “*User Control and Freedom*”, participants highlighted the issue related to the “Back Button”. In “*Flexibility and Efficiency of Use*”, users identified the “Back Button”, the “Lateral Index Navigation”, the “Clickable Elements” and the “Table Filters”. The issues related to “*Visibility of System Status*” are “Breadcrumbs” and “Lateral Index Navigation”. In the “*Information Coding*” heuristic, users identified the “Category Chart Colors” issue. Finally, regarding “*Spatial Organization*” participants highlighted the “Lateral Index Navigation” issue.

Despite the issues identified during the evaluation process, the overall results of the usability assess-

ment remain favorable. The majority of the identified issues were rated as 2 on the severity scale. This indicates that these issues are considered minimal usability concerns and do not require immediate attention or correction. While it is important to acknowledge and address these issues to further enhance the user experience, they are not considered critical to the functioning of the platform.



# 6

## Conclusion

### Contents

6.1 Final Discussion . . . . .	73
6.2 Current Limitations . . . . .	74
6.3 Future Work . . . . .	77

### 6.1 Final Discussion

*CMU Portugal* is a partnership between *CMU* and several institutions in Portugal. This collaborative initiative aims to foster research, innovation, and education in key areas of technology and engineering. Also, the partnership was established to promote knowledge exchange and joint research projects in Portugal.

To evaluate the impact of an international collaboration, such as *CMU Portugal*, and the online identification of its researchers, it is necessary to define metrics and criteria to assess this impact. With this in mind, the use and evaluation of bibliometric data and research output emerge as a possible solution. This document's solution proposed creating a platform that automates and simplifies this process by using *Google Scholar* to track the research output of *CMU Portugal* and its researchers. By doing this,

we cannot only track *CMU Portugal* research output but also measure the influence, scientific impact, and international partnership caused by this program.

*Google Scholar* ranks all the publications on its repository with its *PageRank* algorithm, which attributes a *PageRank* score to an article based on its citations count and links to other important pages [6]. This attributed score can thus work as a factor to quantify the scientific impact of research output. As a consequence of this algorithm, researchers also get higher online visibility for their contributions through *Google Scholar*'s researcher profile [8]. This data repository was chosen over others due to most of *CMU Portugal*'s publications being conference papers [2], and *Google Scholar* also giving importance to this kind of publication. What's more, its *PageRank* algorithm acceptably reflects the research impact of each publication and researcher, making it a suitable tool for this purpose [6].

In our approach, we extracted information and data from *Google Scholar* of *Ph.D.* and affiliated authors/researchers, as well as publications that are both, part of the *CMU Portugal* program and accessible through *Google Scholar*. We used this gathered data to develop a platform that works as an information dashboard that reflects *CMU Portugal*'s influence over scientific contribution. By extracting this information, we have automated the process of extracting bibliometric data of *Ph.D.* and affiliated students, enabling us to evaluate the impact caused by *CMU Portugal*.

After the platform was developed and operational, we evaluated the dashboard by conducting a *User Research* with 19 users that were familiar with academic output, bibliometric data, and data repositories for scientific documents. The participants engaged in a series of predefined tasks and provided feedback on the platform's usability by using a heuristic evaluation checklist and by identifying and rating usability issues. Despite a few minor issues, the results of the evaluation indicate that the platform has achieved its objective of providing a user-friendly interface for exploring the research output and impact of *CMU Portugal*.

Overall, our automated data extraction process and the subsequent visualization of the extracted data through our platform offer valuable insights and a deeper understanding of the impact generated by *CMU Portugal* in the realm of academic research. This work contributes to the advancement of bibliometric analysis and serves as a resource for evaluating the impact of *CMU Portugal*'s initiatives.

## 6.2 Current Limitations

While our methodology for extracting data from *Google Scholar* and conducting bibliometric analysis has proven to be effective, it is important to acknowledge the limitations that we have encountered. These limitations primarily stem from the nature of the data available on *Google Scholar* and the dependencies we have on them.

### 6.2.1 Author's Exclusion

It is important to note that *Google Scholar* provides a vast amount of academic information. However, the data that we can extract is still limited. Our current approach focuses on extracting data related to *CMU Portugal's* authors and publications, specifically targeting *Ph.D.* and affiliated students who have a *Google Scholar* profile page. This means that we currently do not capture the full spectrum of academic collaboration and impact within *CMU Portugal*. We recognize that there may be valuable contributions from other types of researchers or collaborators that are not represented in our analysis.

### 6.2.2 Publications Outside the Scope of *CMU Portugal*

In our approach, we have used a specific timeframe for including publications in our analysis. We consider publications that fall within the author's affiliation with *CMU Portugal*, including their start and end dates, with the addition of a one-year margin. This approach aims to capture a snapshot of the author's academic output during their time with *CMU Portugal*.

However, it is important to acknowledge that this methodology presents certain challenges and potential limitations. One of the main concerns is the possibility of including publications that are not directly related to *CMU Portugal*. This can occur because, during the one-year margin, the student is no longer officially affiliated with *CMU Portugal*. As a result, some publications within that period might not necessarily represent the research conducted under the scope of *CMU Portugal*. Additionally, there is also the possibility of excluding relevant publications that were published after the one-year margin but are still within the scope of *CMU Portugal*. This issue arises because some publications may be published later than expected.

### 6.2.3 *Google Scholar's* Data Update

Additionally, our methodology relies on the data that is available and updated on *Google Scholar*. While *Google Scholar* strives to maintain an extensive database, it is subject to changes made by both the users and *Google Scholar*. Therefore, the accuracy of some of the extracted data is dependent on the reliability of the data from *Google Scholar* itself. This can introduce certain uncertainties and limitations in our analysis.

#### 6.2.3.A *Current Affiliations Inaccuracies:*

One of the potential sources of data inaccuracies in our methodology is the reliance on the author's affiliation as extracted from their *Google Scholar* profile. It is important to note that this affiliation information is provided by the author themselves and is subject to potential changes over time. As a result, there

is a possibility that the affiliation we extract from the currently available information may not accurately reflect the author's affiliation during their time with *CMU Portugal*.

#### **6.2.3.B Profiles Deletion:**

Another issue is the possibility of users deleting their *Google Scholar* profiles. If a user chooses to delete their profile, we may lose access to their information, including their publications and affiliations, if we have not already extracted it prior to the deletion. This can impact the completeness and accuracy of our dataset, especially if the deleted profiles contain significant contributions related to *CMU Portugal*.

### **6.2.4 International and Inter-institutional Collaboration Identification:**

One of our primary objectives in this project was to identify and analyze international and inter-institutional collaborations within the academic publications associated with *CMU Portugal*. While we made progress in this direction, it is important to acknowledge that we did not fully achieve these goals.

#### **6.2.4.A International Collaboration Identification:**

In terms of international collaboration, our approach focused on identifying the students' advisors listed in a publication. If both advisors were recognized as authors in a document, it indicated the presence of an international collaboration. This method allowed us to partially identify international collaborations within the publications associated with international collaboration. However, it is essential to note that this approach may not cover all instances of international collaboration, as there could be scenarios where collaborations occur without the direct involvement of advisors.

#### **6.2.4.B Inter-institutional Collaboration Identification:**

In the case of inter-institutional collaboration, we encountered some difficulties. The reliability of the affiliation data extracted from the author's profile posed a limitation. The affiliation mentioned in the author's profile may change over time, making it difficult to determine the exact institutional collaborations for each publication accurately. Since a *CMU advisor* and a *PT advisor*, each represent their respective institutions, we can identify inter-institutional collaboration by finding both advisors as authors in a publication. But once again, this may not cover all instances of inter-institutional collaboration. Therefore, our current implementation does not fully capture inter-institutional collaborations within academic publications.



### 6.2.5 External Features

One of the limitations of our current implementation is that there are external features that are not directly integrated into the final platform. These features are running the code for “*Data Extraction*”, “*Data Updating*”, “*Data Crossing*” modules, and “*Delete Publications*” (Section 4.5.5.B). These features can be executed by running an executable file that opens a graphical interface that allows the user to run our implementation’s code and delete publications. These external features require users to perform operations outside the main platform, leading to a disjointed user experience and potential difficulties in navigating between different tools or interfaces.

### 6.2.6 JSON Data Handling

Additionally, another limitation in our implementation is the use of multiple *JSON* files throughout our solution. Relying on them extensively in our implementation introduces inefficiencies and challenges in terms of data management. Firstly, using multiple *JSON* files can lead to scattered and fragmented data, making it difficult to maintain a centralized and organized data structure. It becomes difficult to handle and track data across different files, especially as the volume of data increases. Moreover, interacting with these *JSON* files directly through the platform is hindered by the *CORS* policy. The *CORS* policy restricts web applications from accessing resources on different domains, which prevents seamless integration and interaction with the *JSON* files within the platform.

## 6.3 Future Work

### 6.3.1 ORCID Profiles

To address the limitation related to the author’s current affiliation data obtained from their Google Scholar profile, we propose leveraging the capabilities of the *ORCID* platform. *ORCID*, a non-profit organization, serves as a centralized repository of research-related profiles, offering unique digital identifiers for researchers and their professional information [15].

*ORCID* provides an overview of a researcher’s professional history, including their affiliations over the years (Figure 6.1). This historical perspective allows us to gain insights into the author’s affiliations during their time with *CMU Portugal*.

We could use web scraping techniques to extract data from the researcher’s *ORCID* profile, specifically focusing on the timeframe when they were affiliated with *CMU Portugal*. By retrieving this information, we can determine the author’s precise affiliation during their involvement with *CMU Portugal* and accurately link it to their respective publications.

**Biography**  
 PhD 2005  
 Researcher at INESC-ID  
 Professor at Instituto Superior Técnico / Universidade de Lisboa

**Activities**
Collapse all

Employment (3)
 Sort

Universidade de Lisboa Instituto Superior Técnico: Lisboa, Lisboa, PT

2021 to present | Full Professor  
Employment
Show more detail

Source: Ines Lynce

Universidade de Lisboa Instituto Superior Técnico: Lisboa, PT

2013 to 2021 | Professor Associado (Departamento de Engenharia Informática)  
Employment
Show more detail

Source: Ines Lynce

Figure 6.1: Author's ORCID profile page

### 6.3.2 Documents Affiliation Identification

In order to tackle the limitation of inter-institutional collaboration identification, we could leverage *Natural Language Processing (NLP)* techniques, like *Named Entity Recognition (NER)* [43]. *NER* focuses on extracting and classifying named entities from text, including entities such as organizations, locations, and person names.

By applying this technique to the academic document's text, we can specifically target and extract institutional entities mentioned within the document. These institutional entities might represent affiliations that have contributed to the research. *NER* algorithms are trained on large corpora of text and are capable of recognizing patterns and context to accurately identify relevant entities.

### 6.3.3 Use *CMU Portugal* as Keyword for Searching

To expand the identification of authors and publications that fall within the scope of *CMU Portugal*, we can leverage the capabilities of the *scholarly* library. The *scholarly* library offers a resource that enables users to search for specific keywords and retrieve relevant publications [22]. A viable approach would be to perform a targeted search using "*CMU Portugal*" as a keyword. This search would yield a list of publications, of which the vast majority are affiliated with *CMU Portugal*.

Also, by examining each publication within the search results, we can identify the authors who are associated with *CMU Portugal*. If the authors had a *Google Scholar* profile page, we could extract information about them and enrich our dataset.

#### **6.3.4 Create Database for Extracted Data**

One potential solution to address the limitation of relying on multiple *JSON* files in our implementation is to transition to a database-driven approach. By creating a dedicated database for storing the extracted data, we can overcome the inefficiencies and challenges associated with managing data through *JSON* files. A database offers a more efficient and structured way to organize, query, and manipulate data, allowing for better performance and scalability.

#### **6.3.5 Correct Usability Issues**

In order to enhance the overall user experience and improve the final platform, we would have to address the usability issues that were identified during the evaluation in *Section 5.4*. Resolving these issues helps ensure that users can navigate and interact with the platform effortlessly and efficiently.

By acknowledging and addressing these usability issues, we can make the necessary adjustments and refinements to optimize the platform's functionality and user interface.



# Bibliography

- [1] H. Horta and M. T. Patrício, “Setting-up an international science partnership program: a case study between portuguese and us research universities,” *Technological Forecasting and Social Change*, vol. 113, pp. 230–239, 2016.
- [2] CMU Portugal, “CMU Portugal 2018/2019 Anual Report,” visited on 2022-03-22. [Online]. Available: [https://www.cmuportugal.org/wp-content/uploads/2020/09/Relatorio\\_CMU.pdf](https://www.cmuportugal.org/wp-content/uploads/2020/09/Relatorio_CMU.pdf)
- [3] S. M. Pfotenhauer, J. S. Jacobs, J. A. Pertuze, D. J. Newman, and D. T. Roos, “Seeding change through international university partnerships: The mit-portugal program as a driver of internationalization, networking, and innovation,” *Higher Education Policy*, vol. 26, no. 2, pp. 217–242, 2013.
- [4] M. T. Patrício, P. Santos, P. M. Loureiro, and H. Horta, “Faculty-exchange programs promoting change: motivations, experiences, and influence of participants in the carnegie mellon university-portugal faculty exchange program,” *Tertiary Education and Management*, vol. 24, no. 1, pp. 1–18, 2018.
- [5] M. D. Hird and S. M. Pfotenhauer, “How complex international partnerships shape domestic research clusters: Difference-in-difference network formation and research re-orientation in the mit portugal program,” *Research Policy*, vol. 46, no. 3, pp. 557–572, 2017.
- [6] J. van Aalst, “Using google scholar to estimate the impact of journal articles in education,” *Educational researcher*, vol. 39, no. 5, pp. 387–400, 2010.
- [7] A. Diem and S. C. Wolter, “The use of bibliometrics to measure research performance in education sciences,” *Research in higher education*, vol. 54, no. 1, pp. 86–114, 2013.
- [8] C. Rovira, L. Codina, F. Guerrero-Solé, and C. Lopezosa, “Ranking by relevance and citation counts, a comparative study: Google scholar, microsoft academic, wos and scopus,” *Future Internet*, vol. 11, no. 9, p. 202, 2019.
- [9] B. Thoma and T. M. Chan, “Using google scholar to track the scholarly output of research groups,” *Perspectives on medical education*, vol. 8, no. 3, pp. 201–205, 2019.

- [10] UT Austin, "UT Austin Portugal," visited on 2022-03-22. [Online]. Available: <https://utaustinportugal.org/>
- [11] MIT, "MIT Portugal," visited on 2022-03-22. [Online]. Available: <https://www.mitportugal.org/>
- [12] P. Mongeon and A. Paul-Hus, "The journal coverage of web of science and scopus: a comparative analysis," *Scientometrics*, vol. 106, no. 1, pp. 213–228, 2016.
- [13] Clarivate, "Web of Science," visited on 2022-05-13. [Online]. Available: <https://clarivate.com/webofsciencigroup/solutions/web-of-science/>
- [14] Scopus, "Scopus," visited on 2022-05-14. [Online]. Available: [https://service.elsevier.com/app/answers/detail/a\\_id/15534/supporthub/scopus/#tips/](https://service.elsevier.com/app/answers/detail/a_id/15534/supporthub/scopus/#tips/)
- [15] ORCID, "ORCID," visited on 2022-05-10. [Online]. Available: <https://orcid.org/>
- [16] B. Zhao, "Web scraping," *Encyclopedia of big data*, pp. 1–3, 2017.
- [17] IBM, "IBM," visited on 2023-05-08. [Online]. Available: <https://www.ibm.com/topics/api/>
- [18] SerpAPI, "Google Scholar SerpAPI," visited on 2022-05-06. [Online]. Available: <https://serpapi.com/google-scholar-api>
- [19] ScraperAPI, "ScraperAPI," visited on 2022-05-12. [Online]. Available: <https://www.scraperapi.com/blog/best-google-scholar-apis-proxies/>
- [20] ScraperAPI, "ScraperAPI Proxy API," visited on 2022-05-12. [Online]. Available: <https://www.scraperapi.com/>
- [21] D. Kouzis-Loukas, *Learning Scrapy*. Packt Publishing Ltd, 2016.
- [22] pypi, "scholarly," visited on 2023-05-1. [Online]. Available: <https://pypi.org/project/scholarly/>
- [23] S. A. Cholewiak, P. Ipeirotis, V. Silva, and A. Kannawadi, "SCHOLARLY: Simple access to Google Scholar authors and citation using Python," 2021. [Online]. Available: <https://github.com/scholarly-python-package/scholarly>
- [24] Microsoft, "Microsoft Academic," visited on 2022-05-06. [Online]. Available: <https://www.microsoft.com/en-us/research/project/academic/>
- [25] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.
- [26] M. Sysel and O. Doležal, "An educational http proxy server," *Procedia Engineering*, vol. 69, pp. 128–132, 2014.

- [27] ScraperAPI, “ScraperAPI,” visited on 2023-01-10. [Online]. Available: <https://www.scraperapi.com/blog/10-tips-for-web-scraping/>
- [28] D. Perino, M. Varvello, and C. Soriente, “Proxytorrent: Untangling the free http (s) proxy ecosystem,” in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 197–206.
- [29] Outsystems, “Outsystems,” visited on 2023-02-7. [Online]. Available: <https://www.outsystems.com/>
- [30] Microsoft, “Power BI,” visited on 2023-02-5. [Online]. Available: <https://powerbi.microsoft.com/pt-pt/>
- [31] D3.js, “D3.js,” visited on 2022-11-30. [Online]. Available: <https://d3js.org/>
- [32] Microsoft, “Microsoft Visual C++ Build Tools,” visited on 2022-11-30. [Online]. Available: <https://visualstudio.microsoft.com/pt-br/downloads/>
- [33] PyPi, “lxml,” visited on 2022-11-30. [Online]. Available: <https://pypi.org/project/lxml/>
- [34] —, “XlsxWriter,” visited on 2023-02-5. [Online]. Available: <https://pypi.org/project/XlsxWriter/>
- [35] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.
- [36] Dev, “Build Your Own Google Scholar API With Python Scrapy,” visited on 2022-05-12. [Online]. Available: <https://dev.to/iankerins/build-your-own-google-scholar-api-with-python-scrapy-4p73>
- [37] L. Engqvist and J. G. Frommen, “The h-index and self-citations,” *Trends in ecology & evolution*, vol. 23, no. 5, pp. 250–252, 2008.
- [38] J. Chen, J. Jiang, H.-X. Duan, T. Wan, S. Chen, V. Paxson, and M. Yang, “We still don’t have secure cross-domain requests: an empirical study of cors.” in *USENIX Security Symposium*, 2018, pp. 1079–1093.
- [39] F. Lundh, “An introduction to tkinter,” *URL: www.pythonware.com/library/tkinter/introduction/index.htm*, 1999.
- [40] D. Dowding and J. A. Merrill, “The development of heuristics for evaluation of dashboard visualizations,” *Applied clinical informatics*, vol. 9, no. 03, pp. 511–518, 2018.
- [41] Nielsen Norman Group, “Powers of 10: Time Scales in User Experience,” visited on 2023-04-15. [Online]. Available: [https://www.nngroup.com/articles/powers-of-10-time-scales-in-ux/?fbclid=IwAR0qsOSKooC3wN98YY1RUL5p2Qww\\_DviAyBeOsdyl1wuouwdBidaFDv7-4w](https://www.nngroup.com/articles/powers-of-10-time-scales-in-ux/?fbclid=IwAR0qsOSKooC3wN98YY1RUL5p2Qww_DviAyBeOsdyl1wuouwdBidaFDv7-4w)
- [42] D. G. Manuel J. Fonseca, Pedro Campos, *Introdução ao Design de Interfaces*. FCA, 2017.
- [43] A. Bodnari, L. Deleger, T. Lavergne, A. Neveol, and P. Zweigenbaum, “A supervised named-entity extraction system for medical text.” in *CLEF (Working Notes)*, 2013.







# Demographic Questionnaire

## A.1 Preparation

User tests will be developed in order to evaluate the platform developed for our master's thesis dissertation. These tests will be performed with several users that are familiar with academic research, bibliometric data, and search engines for academic literature. The necessary equipment to perform this evaluation is a computer with access to the internet, a browser, and access to the developed platform. Furthermore, we will be asked to the users to answer to 2 questionnaires, one performed before the testing in order to gather demographic data, and one after the testing with the objective to evaluate the final appreciation of the users regarding the platform. We will serve as test observants and coordinators, taking notes when necessary.

## A.2 Questionnaire

1. *I accept that demographic data and information are collected anonymously as a means of evaluating the dashboard and for statistical purposes.*

- Yes
- No

2. *Sex:*

- Male
- Female
- Rather Not Answer

3. *Age Group:*

- 18 - 21
- 22 - 30
- 31 - 40
- > 40

4. *What is the highest school degree you have completed?*

- High School
- Bachelor's degree
- Master's degree
- Ph.D.

5. *Have you ever searched for academic literature?*

- Yes
- No

6. *Have you ever used a search engine for academic literature?*

- Yes
- No

7. *How would you classify your level of experience with technological devices?*

- Inexperienced
- Familiarized
- Experienced

8. *How easy is it for you to learn a new technologic device or application, by exploring, without help?*

- Very Easy
- Easy
- Normal

- Hard
- Very Hard





# User Guide

## B.1 Introduction

The platform was developed as a dashboard with the objective of visualizing and accessing the bibliometric data of several Ph.D. and Affiliated Students of the CMU Portugal Program. With the performed tests we intend to evaluate the effectiveness, efficiency, and overall satisfaction of this platform, as well as to identify possible problems with the platform.

During the execution of the tests, the users will be asked to perform 3 tasks regarding the platform and to answer a final questionnaire. Each task will be timed. As referenced before, the purpose of these user tests is to identify future improvements to the dashboard and being the platform the only element that is being evaluated. For this reason, the user should not concern about how much time it takes him to complete the tasks, as well as be worried about any difficulty found or error made while trying to complete these tasks.

Any doubt from the user must be clarified before the beginning of testing, being that during the tests the users should not ask for help for results-gathering efficiency reasons. The user is free to abandon the test at any point of the execution if he wishes to. These tests have an estimated duration of 15 to 20 minutes. We thank you in advance for your help and collaboration.

## B.2 Evaluation

### 1. *Global Dashboard:*

1. Identify the total number of citations (tell the evaluator).
2. Verify how many citations there were in 2021 (tell the evaluator).
3. Verify how many publications of the type “article” were published in 2016 (tell the evaluator).
4. Without leaving the current page, sort the list of publications by “Citations Count”.
5. Without leaving the current page, sort the list of authors by “Number of Publications”.
6. Select the author in the third position.
7. Return to the global dashboard.
8. Decrease the horizontal size of the navigation tab and increase the horizontal size of the dashboard window, with the same action.
9. Download the displayed data.

### 2. *Authors:*

1. Navigate to the list of authors.
2. Sort the list of authors by Name.
3. Find the student “Dragana Bajovic”.
4. Identify the type of the student (tell the evaluator).
5. Navigate to the student’s profile page.
6. Verify the student’s CMU Advisor (tell the evaluator).
7. Without leaving the current page, sort the list of publications by “Citations Count”.
8. Select the first publication that appears with only “Just CMU Advisor” as an international collaboration.
9. Return to the previous page.
10. Without leaving the current page, identify which students have collaborated with the student “Dragana Bajovic” (tell the evaluator).
11. Select the student in the first position.
12. Return to the total list of authors.

3. *Publications:*

1. Navigate to the list of publications.
2. Sort the list of publications by “Student Collaboration”.
3. Select the publication with 50 citations.
4. Identify the title of the publication (tell the evaluator).
5. Identify how many students collaborated on this publication (tell the evaluator).
6. Without leaving the current page, sort the list of authors by “Citations Count”.
7. Select the student with the name “Dusan Jakovetic”.
8. Return to the total list of publications.
9. Navigate to the Global Dashboard.







# Usability Questionnaire

## C.1 Final Balance

### C.1.0.A Used Heuristics

- **Visibility of system status:** The system should always keep the user informed about what is going on through appropriate feedback within a reasonable time.
- **Match between system and the real world:** The system should speak the user's language, with words, phrases, and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
- **User control and freedom:** Users should be free to select and sequence tasks (when appropriate), rather than having the system do this for them. Users will need a clearly marked “emergency exit” to leave the unwanted state without having to go through an extended dialogue. Users should make their own decisions regarding the costs of exiting their current work.
- **Consistency and standards:** Users should not have to wonder whether different words, situations, or actions mean the same thing.
- **Recognition rather than recall:** Make objects, actions, and options visible. The user should not

have to remember information from one part of the dialogue to another. Instructions for the use of the system should be visible or easily retrievable whenever appropriate.

- **Flexibility and efficiency of use:** The system should offer users several options when it comes to finding content. Users should be able to achieve their goals in an efficient manner.
- **Aesthetic and minimalist design:** Dialogues should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
- **Spatial organization:** Relates to the overall layout of a visual representation and refers to how easy it is to locate an information element in the display and the distribution of elements in representations.
- **Information coding:** Refers to the use of symbols or representations to aid perception.
- **Orientation:** Provision of support for the user and help to orientate them in the visualization.

#### C.1.0.B Rating

1. *For each of the executed tasks, choose a number between 1 and 5 (1 being totally disagreed and 5 being totally agreed), in order to identify if you agree or not with that our platform respected the previous heuristics list:*

**Table C.1:** Heuristics Rating

Heuristic	Task 1	Task2	Task3
Visibility of system status			
Match between system and the real world			
User control and freedom			
Consistency and standards			
Recognition rather than recall			
Flexibility and efficiency of use			
Aesthetic and minimalist design			
Spatial organization			
Information coding			
Orientation			

**2. Is there any aspect in particular of the platform that you did not enjoy?**

- Yes
- No

***If you answered yes on the previous question, which aspects did you not enjoy? Classify them according to the following scale (1 - 4):***

1. Aesthetic problem only (Does not need to be corrected).
2. Minor usability problem (Can be corrected, but it is not urgent).
3. Major usability problem (It is important to be corrected).
4. Usability Catastrophe (It is imperative to be corrected).

**3. Other observations that you wish to point out:**

***Thank you for your collaboration!***