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

João Antunes
INESC-ID
Instituto Superior Técnico
Lisbon, Portugal
joao.c.jeronimo.antunes@tecnico.ulisboa.pt

ABSTRACT

The Carnegie Mellon University Portugal (*CMU Portugal*) program is an international collaboration between Carnegie Mellon University (*CMU*) and several Portuguese institutions. This program aims to put Portugal at the forefront of technological advancements by promoting education, research, innovation, and institutional 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, 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 *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

CMU Portugal, International Partnership, Data Extraction, Bibliometric Data, *Google Scholar*, Research Impact

1 INTRODUCTION

1.1 Problem Definition

Since 2006, there have been several international research and innovation collaborations between institutions from the United States and Portuguese organizations and universities. One of these collaborations is *CMU Portugal* [8] 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 to assess the quality of the resulting research output. Several studies ([13] [12] [7] [16] [4] [14] [15]) 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 [16].

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 is to analyze *CMU Portugal's* bibliometric data.

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. 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.

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 *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 *CMU* and *PT* (Portugal) advisor, where each advisor represents its respective institution. Because of this, we can identify international and inter-institutional 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. Through various charts, graphs, and tables, users can gain insights into publication trends, author profiles, collaboration networks, and research impact.

1.4 Document Organization

This thesis is structured as follows: First, *Section 2* gives an introduction to what the *CMU Portugal* program consists of. Then, *Section 3* discusses related work, which includes studies on how data repositories can be used as a tool to access scientific impact. The proposed solution and its implementation are presented in *Section 4* 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. *Section 6*, presents a summary of this thesis's most important key features and the current limitations of our implementation.

2 *CMU PORTUGAL*

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. Currently, the major objective is "to bring up interdisciplinary collaboration between industry and academia across different levels of 'big data' development stack" [2]. To achieve this objective, *CMU Portugal* has several initiatives and programs [2]. A few of these initiatives will be described in the following sections:

2.1 Talent Development

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 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*.

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*), as well as 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* projects, these aim to foster new initiatives and promote information and communication technologies projects and integrative research in strategic emerging areas.

3 RELATED WORK

To facilitate the exploration and evaluation of academic research, there are data repositories that store scholarly output and bibliometric data. These repositories provide researchers with a platform to showcase their work. Within these repositories, users are often assigned unique profiles that aggregate their publications, citation counts, *h-indexes*, and other bibliometric indicators, where researchers can gain insights into their research outputs [15].

3.1 *Scientific and Academic Research Data Repositories:*

The article entitled *The use of bibliometrics to measure research performance in education sciences* [4] article investigates the performance and impact of educational research professors through the use of bibliometric data from *Google Scholar* and *WoS* platforms while doing a comparison between the two. This study [4] 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.

3.1.1 ***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 the publication's titles, authors, year, source information, redirecting links to full-text documents, citation count, a list of citing documents, and hyperlinks to these documents [16]. This platform features a *PageRank* algorithm to sort the included publications [16]. This algorithm is explained in more detail in *Section 3.3.2*.

3.1.2 ***Web of Science and Scopus:*** *Web of Science (WoS)* and *Scopus* are scientific journal search and indexing databases. In particular, *WoS* is considered to be the "world's most trusted publisher-independent global citation database" [1]. Both these platforms

feature journals and articles that are classified by field of research, country, and language [10].

3.2 *Research Output Impact Assessment:*

The conducted study in the *Ranking by relevance and citation counts, a comparative study: Google Scholar, Microsoft Academic, WoS, and Scopus* [14] 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" [14].

The conclusion of this document [14] states that both *Google Scholar* and *Microsoft Academic* rely mostly upon their citations 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.

3.3 *Tracking Scholarly Output through Google Scholar:*

3.3.1 ***Using Google Scholar to Track the Scholarly Output of Research Groups.*** In the *Using Google Scholar to track the scholarly output of research groups* publication [15], 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" [15] 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 [16], this profile makes it possible to rank authors according to their citations count and the *h-index* in which the first h articles from the Author's documents 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" [14].

By the end of this article [15], it is concluded that *Google Scholar* provides an efficient and scalable approach to tracking the scholarly output of each research group.

3.3.2 ***Google Scholar Evaluation of Journal Articles Impact in Education.*** The *Using Google Scholar to estimate the impact of journal articles in education:* document [16] 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, but not entirely, on the citation count of each publication [14], "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" [16]. 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 its *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 [16].

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 [16] 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 subfields. Finally, the great majority of citations from *Google Scholar* were from peer-reviewed documents.

4 IMPLEMENTATION

In this solution, we extracted the *Google Scholar's* available bibliometric data, regarding its researchers and publications. 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 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.

Google Scholar was chosen for the following reasons:

- A big majority of the *CMU Portugal's* publications are conference papers [2], therefore, these have a 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 [16].
- *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 [16].
- The previously referred studies demonstrated that *Google Scholar's* citation count-based algorithm and researcher's profile are acceptable and accurate tools for assessing an article's and author's scientific contribution and importance [16] [15].

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* and 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. 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. The number of student collaborations can also be 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 stopped. 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.

4.1.2 *Publications:*

(1) **Authors:** We require the list of authors to identify international collaboration between researchers and 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 research area 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"* (*"incollection"*), *"Pre-prints"*, *"Abstracts"*, *"Technical Reports"*, and other scholarly literature.

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.

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 *Excel* file. 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*.

4.2.2 ***Google Scholar's Data:*** 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.

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 *PDF* file, and the number of citations per year. 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"

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, this field is labeled as *"Source"* and it does not provide explicit information about the publication type.

4.3 System Architecture

Our system architecture is designed with a modular approach, consisting of four key modules: *Data Scraping*, *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 1*.

- ***Data Scraping*:** The *Data Scraping* serves as the foundation, responsible for extracting relevant information from *Google Scholar*.
- ***Data Update*:** The Data Update module plays a crucial role in ensuring that our data is up-to-date.
- ***Data Crossing*:** This module receives information about authors and their publications. This data is used to extract valuable insights by crossing and evaluating various data points and fields.
- ***Final Platform*:** The *Final Platform* is a user-friendly dashboard to visualize the extracted data. The platform allows for exploring authors' profiles, publications' information, citation analysis, and more. providing a view of the academic impact caused by *CMU Portugal*.

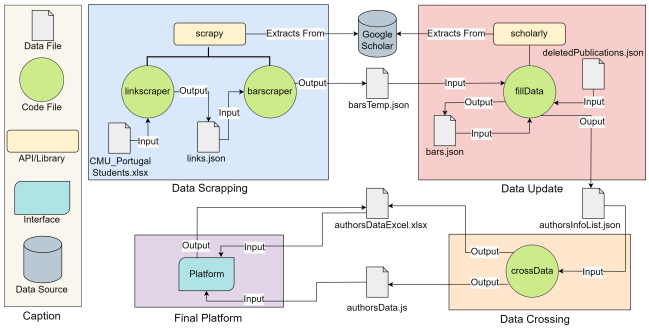

Figure 1: *Information Gathering Architecture*

4.4 Methodology

This implementation uses the *Python3* programming language [17] with the *scholarly*, and *scrapy* libraries/*APIs* to extract data from *Google Scholar*.

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.

4.4.1 ***Data Scraping:*** This module contains two *Python* code files: *linkspider.py and barspider.py*. Each one of these code files (*spiders*) is an agent that can "crawl" along an *HTML* page and extract the necessary information [3].

4.4.1.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. 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.

In order to identify which publications are under the scope of *CMU Portugal*, we only considered publications published under the author's *Start Research* year and the *End Research* year, plus one year as a margin.

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"* where each entry represents a publication and its respective information.

The information on each publication is separated by fields:

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

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

4.4.1.B ***barspider:*** This spider receives as input the *"links.json"* file generated by 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. From the document's *Google Scholar* page, 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"* (Figure 2). This file shares a similar structure to the *"links.json"* data file, and each entry represents a publication.
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.
- ***pub_type*:** The type of the 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.


```
{
  "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 2: *"barsTemp.json" Publication Entry*

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.

4.4.2 ***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. The update process for this file consists of three distinct stages:

- ***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"*. This can happen when authors delete their *Google Scholar* profile page before the last data update.
- ***Filter Deleted Publications*:** Checks if any publications listed in *"barsTemp.json"* were previously deleted by the user. 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." This can happen because we are only able to identify three types of publications directly from *Google Scholar*. 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. 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. At the end of this step, the information is moved into a file called *"authorsInfoList.json"*.

4.4.2.A ***Structure Data:*** This step is used to reorganize *"bars.json"*. 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.

Given this, the information in *"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.

4.4.3 ***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. By leveraging this data, we initiate a series of data cross-referencing operations to extract valuable insights and generate new information for visualization purposes on the final platform.

The *Data Crossing* module is structured into several sequential steps. Firstly, we cross-reference the information about the authors. Next, we proceed to cross-reference the publications' data.

Lastly, we employ the collective information from all authors and publications to generate a "profile" that represents *CMU Portugal* as a whole.

4.4.3.A ***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 publications. 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. To address this anomaly, we implemented a filtering mechanism to discard any citations that occurred before the year of publication.

Once this problem was removed, we were able to calculate the following data:

- ***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.
- ***h-index*:** The *h-index* represents the number of an author's articles (*h*) that have garnered a minimum of h citations [6].
- ***h-index (Last 5 Years)*:** The *h-index* but of publications that were published in the last 5 years.
- ***i10-index*:** This index has the same formula as the *h-index*, but *h* stands 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.

4.4.3.B ***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. Another analysis was conducted to count the number of publications per type and the number of publications per research area for each author.

4.4.3.C ***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. 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.

4.4.3.D ***Calculate the Number of Collaborations Between Students:*** In addition to evaluating collaborations between advisors and students, we also considered the 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. 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.

4.4.3.E ***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.

4.4.3.F ***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.

Within the *"authorsData.js"* file, we organized the collected information under the *"CMUPortugal"* field. 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*. Each individual author within this field contains their respective information and list of publications.

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*. 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.

4.4.4 ***Final Platform:*** Once the *"authorsData.js"* file was complete and contained all the relevant information, we proceeded to implement our platform. In our methodology, we used *HTML5*, *CSS*, and *JavaScript*, along with the *D3* library, to develop the final platform. The *D3* library was used to create the generated charts.

The platform was designed with three functionalities: the *"Global Dashboard"*, *"Authors"*, and *"Publications"*. 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.

4.4.5.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*. 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 (*Figure 3* and 4).

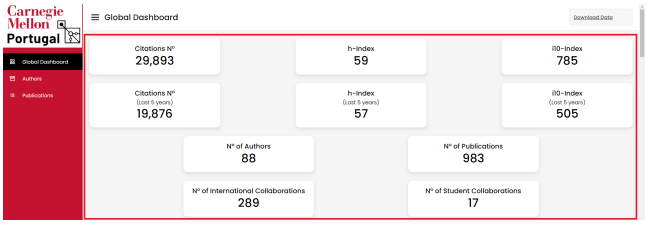

Figure 3: ***Global Dashboard's Quantitative Metrics***

4.4.5.B ***Authors:*** The *"Authors"* section of the platform allows users to delve into individual author profiles.

- ***Authors List:*** In this functionality, the initial page presents a list of authors associated with *CMU Portugal*. This list provides an overview of all the students. Each student is represented by a card containing the author's picture and partial information.
- ***Author's Profile:*** By clicking on an author entry in the list of authors, users gain access to the author's profile. This profile presents the 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*.

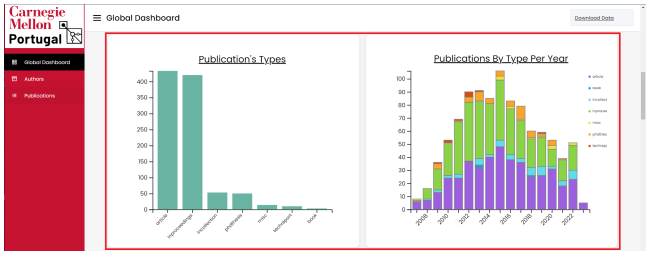

Figure 4: ***Global Dashboard's Charts***

Furthermore, the profile features a table that presents a list of the author's publications. Each entry in the table is clickable, leading to the publication's profile page. Additionally, the author's profile includes a table that lists other students who have collaborated with the author.

4.4.5.C ***Publications:*** The *"Publications"* functionality focuses on providing detailed information about each publication associated with *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*.
- ***Publication's Profile:*** By clicking on an entry in the table, users can access the publication's profile. Similar to the authors' profiles, the publication's profile includes an information "card" that contains specific details about the publication sourced from *Google Scholar*. This profile also presents quantitative information and a table listing the students who contributed to the publication.

5 EVALUATION

For our evaluation process, we adopted the evaluation method outlined in the *"The Development of Heuristics for Evaluation of Dashboard Visualizations"* article [5].

The evaluation process consisted of three stages. First, we created a questionnaire to collect demographic data from the users participating in the evaluation. Next, the users were guided through a series of pre-defined tasks that aimed to familiarize them with the platform's functionalities and features.

Finally, we asked users to answer a usability questionnaire incorporating a heuristic checklist to evaluate their satisfaction level with the platform and identify any areas that required improvement.

5.1 User Characterization

During this stage, we provided 19 participants with a questionnaire to collect demographic data. 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*.

5.2 Platform Validation

During the evaluation process, we provided each participant with a user guide that outlined three specific tasks to be performed on the final platform. 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 [11].

Once the participants completed the tasks, we asked them to answer a usability questionnaire. 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 inform the user of what is happening.
(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.
(3) ***User Control and Freedom*:** Users should decide for themselves how much it will cost to stop doing something or exit an undesirable state.
(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.
(6) ***Flexibility and Efficiency of Use*:** When it comes to choosing how to discover content, the system should give users many possibilities.
(7) ***Aesthetic and Minimalist Design*:** Information that is unnecessary or rarely used shouldn't be included in dialogues.
(8) ***Spatial Organization*:** Relates to how a visual representation is organized overall.
(9) ***Information Coding*:** The use symbols or representations to facilitate perception.
(10) ***Orientation*:** Providing assistance to the user and guiding them in the visualization.

We then 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 [9]:

- ***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

As previously mentioned, we recorded the time taken by each user to complete each task. By capturing this data. The recorded times allowed us to calculate the average time spent by users on each task and overall.

5.4 Usability Questionnaire

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.

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 greater 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. On average, this issue received a rating of 2.
- ***Lateral Index Navigation*:** 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. 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. This issue received a rating of 2 based on 5 answers.
- ***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

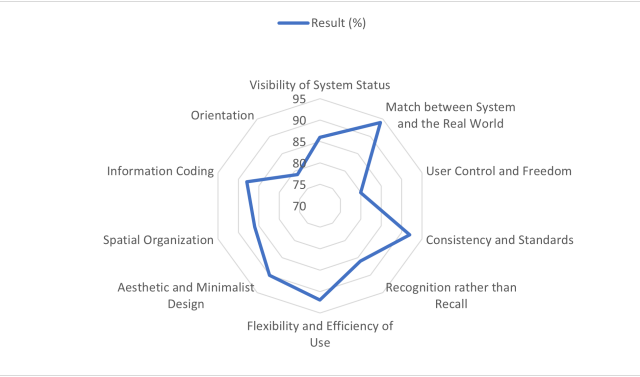

Figure 5: *Heuristic Evaluation Results*

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 to explore the entire dashboard was within a reasonable timeframe (10 minutes and 8 seconds).

Regarding the usability questionnaire results and *Figure 5*, our main issues were related to *"Orientation"*, *"User Control and Freedom"*, *"Flexibility and Efficiency of Use"*, *"Visibility of System Status"* 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 "Table Filters". The issues related to *"Visibility of System Status"* are "Breadcrumbs" and "Lateral Index Navigation". Finally, regarding *"Spatial Organization"* participants highlighted the "Lateral Index Navigation" issue.

Also, the majority of the identified issues were rated as 2 on the severity scale. This indicates that these issues do not require immediate attention or correction.

6 CONCLUSION

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.

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.

In our approach, we extracted information and data 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, 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 familiarized with academic output, bibliometric data, and data repositories for scientific documents. The results of this research indicate that the platform has achieved its objective of providing an interface for exploring the research output and impact of *CMU Portugal*.

6.2 Current Limitations

6.2.1 ***Author's Exclusion:*** Our current approach targets *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*.

6.2.2 ***Publications Outside the Scope of CMU Portugal:*** We have used a specific timeframe for including publications in our analysis. However, there 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*.

ACKNOWLEDGMENTS

This work is co-financed by *Fundação para a Ciência e a Tecnologia* (Portuguese Foundation for Science and Technology) under project *UIDB/50021/2020* and through the *Carnegie Mellon Portugal* program.

REFERENCES

[1] Clarivate. [n.d.]. Web of Science. https://clarivate.com/webofsciencegroup/solutions/web-of-science/ visited on 2022-05-13.
[2] CMU Portugal. [n.d.]. CMU Portugal 2018/2019 Anual Report. https://www.cmuportugal.org/wp-content/uploads/2020/09/Relatorio_CMU.pdf visited on 2022-03-22.
[3] Dev. [n.d.]. Build Your Own Google Scholar API With Python Scrapy. https://dev.to/iankerins/build-your-own-google-scholar-api-with-python-scrapy-4p73 visited on 2022-05-12.
[4] Andrea Diem and Stefan C Wolter. 2013. The use of bibliometrics to measure research performance in education sciences. *Research in higher education* 54, 1 (2013), 86–114. https://doi.org/10.1007/s11162-012-9264-5
[5] Dawn Dowding and Jacqueline A Merrill. 2018. The development of heuristics for evaluation of dashboard visualizations. *Applied clinical informatics* 9, 03 (2018), 511–518.
[6] Leif Engqvist and Joachim G Frommen. 2008. The h-index and self-citations. *Trends in ecology & evolution* 23, 5 (2008), 250–252.
[7] Mackenzie D Hird and Sebastian M Pfotenhauer. 2017. How complex international partnerships shape domestic research clusters: Difference-in-difference network formation and research re-orientation in the MIT Portugal Program. *Research Policy* 46, 3 (2017), 557–572. https://doi.org/10.1016/j.respol.2016.10.008
[8] Hugo Horta and Maria Teresa Patrício. 2016. Setting-up an international science partnership program: a case study between Portuguese and US research universities. *Technological Forecasting and Social Change* 113 (2016), 230–239. https://doi.org/10.1016/j.techfore.2015.07.027
[9] Daniel Gonçalves Manuel J. Fonseca, Pedro Campos. 2017. *Introdução ao Design de Interfaces.* FCA.
[10] Philippe Mongeon and Adèle Paul-Hus. 2016. The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics* 106, 1 (2016), 213–228. https://doi.org/10.1007/s11192-015-1765-5
[11] Nielsen Norman Group. [n.d.]. Powers of 10: Time Scales in User Experience. https://www.nngroup.com/articles/powers-of-10-time-scales-in-ux/?fbclid=IwAR0qsOSKooC3wN98YY1RUL5p2Qww_DviAyBeOsdy1lwuouwdBidaFDv7-4w visited on 2023-04-15.
[12] Maria Teresa Patrício, Patrícia Santos, Paulo Maia Loureiro, and Hugo Horta. 2018. Faculty-exchange programs promoting change: motivations, experiences, and influence of participants in the Carnegie Mellon University-Portugal Faculty Exchange Program. *Tertiary Education and Management* 24, 1 (2018), 1–18. https://doi.org/10.1080/13583883.2017.1305440
[13] Sebastian M Pfotenhauer, Joshua S Jacobs, Julio A Pertuze, Dava J Newman, and Daniel T Roos. 2013. Seeding change through international university partnerships: The MIT-Portugal program as a driver of internationalization, networking, and innovation. *Higher Education Policy* 26, 2 (2013), 217–242. https://doi.org/10.1057/hep.2012.28
[14] Cristòfol Rovira, Lluís Codina, Frederic Guerrero-Solé, and Carlos Lopezosa. 2019. Ranking by relevance and citation counts, a comparative study: Google Scholar, Microsoft Academic, WoS and Scopus. *Future Internet* 11, 9 (2019), 202. https://doi.org/10.3390/fi11090202
[15] Brent Thoma and Teresa M Chan. 2019. Using Google Scholar to track the scholarly output of research groups. *Perspectives on medical education* 8, 3 (2019), 201–205. https://doi.org/10.1007/s40037-019-0515-4
[16] Jan van Aalst. 2010. Using Google Scholar to estimate the impact of journal articles in education. *Educational researcher* 39, 5 (2010), 387–400. https://doi.org/10.3102/0013189X10371120
[17] Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA.