TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

Open Journal Citation Ranking

Master Thesis

Akshaya Dharmaraj

May 8, 2023

Rhineland-Palatinate Technical University of Kaiserslautern-Landau,
Department of Computer Science,
67653 Kaiserslautern,
Germany

Supervisor: Prof. Dr Sebastian Vollmer
Dr David Antony Selby

Abstract

Considering the vast number of papers published daily and the shortcomings of 'subjective' methods, such as peer review, citation analysis presents an 'objective' alternative for research evaluation. When analysing citation data, a crucial aspect to consider is the selection of appropriate data sources. A detailed study of citation data sources and various methods of data extraction is presented. Until recently, many of the commonly used citation databases were proprietary. However, the emergence of new open data sources is transforming the research landscape. Moreover, the studies conducted on some of the more recently established data sources, such as Semantic Scholar and Open Alex, are comparatively few in number.

While a variety of tools are available for extracting data from citation databases, a majority of these tools lack the ability to extract data from multiple databases. As a solution to this problem, we propose the tool "CitaTrack", which provides a uniform interface for extracting and aggregating citation data from multiple databases. To create this tool, five different data sources were utilised, including recently established ones, such as Open Alex and Semantic Scholar, along with Crossref, Open Citations Corpus, and Scopus. The sources were selected based on their multidisciplinary nature, as this enables the extraction of citation data across a wide range of research fields and topics.

An analysis was conducted to rank 20 journals in the field of economics using the Stigler model based on the citation data extracted using the tool. This serves as an illustration of the potential application of the tool in the field of bibliometrics. There is a general consensus regarding the top 5 journals in the field of economics, and this consensus was used as a benchmark to assess and compare the obtained rankings. The rankings were mostly in agreement with the consensus. The ranking generated by Scopus listed all of the top journals among the top 5, while the rankings from the other data sources had four out of the top 5 journals ranked in the top 4 positions.

Acknowledgement

I would like to express my deep gratitude to my supervisor Prof. Dr Sebastian Vollmer, for his invaluable guidance and support throughout the entire dissertation process. I am indebted to my supervisor, Dr David Antony Selby, who has been an excellent mentor, providing insightful feedback, challenging me to think critically, and guiding me to achieve excellence in all aspects of my work.

I would also like to thank my father for his endless encouragement, love, and support. His unwavering belief in my abilities has been a constant source of motivation throughout my academic journey, and I am forever grateful for his support. Finally, I would like to express my gratitude to my dear friends for their encouragement throughout this challenging and rewarding journey.

Contents

List of Figures

List of Tables

1. Introduction

In the ever-expanding world of information science, more and more scientific papers are being written and published by scholars from different domains. In evaluating scientific performance, publications are considered to be an effective indicator due to their ability to capture the many different facets of research (Retzer & Jurasinski, 2009). The publication records of researchers are, therefore, the primary means for assessing their scientific performance. However, it is not an easy task to identify papers of excellence from these innumerable works. The question is: how can one measure the impact of a paper in a particular domain? What are the various tools available for research assessment, and how effective are they in measuring the influence of research works and providing a reliable result?

Traditionally, the research output was evaluated using 'subjective' methods like peer reviews. However, owing to limited resources and increased bureaucratisation in science, peer review is progressively being replaced or complemented with bibliometric methods (Haustein & Larivière, 2014). Moreover, information technology has helped the field of bibliometrics to evolve as an alternative solution to the questions raised above to some extent. Bibliometric analysis can be used to quantify the impact of scholarly works or journals to a certain degree. The field of bibliometrics includes a broad range of techniques and methods such as citation analysis, bibliographical coupling, co-citation analysis, and coauthor analysis (Zupic & Čater, 2015). Citation analysis is one of the main methods used in bibliometrics that studies the relationship between a scientific work and its bibliographic references (the citations) (Debackere et al., 2002). In simple terms, it can be defined as the examination of downstream citation frequency and pattern (Cooper, 2015). In this context, downstream refers to articles produced (in the future) that are influenced by the cited (source) article.

Citation analysis rests on the notion that when a work is referenced in a subsequent paper, it indicates the influence of the cited work on the paper that cites it (Haustein & Larivière, 2014). The citation indexes or databases are the primary resources used for citation analysis. Traditionally, the indexes created by Thomson Reuters have been the main source of citation data used for this purpose (Neuhaus & Daniel, 2008). However, the organisation is no longer the exclusive provider of citation databases as there are other providers like Google Scholar and Scopus now available in the market. Some citation databases are free to use, while others require paid subscriptions. To ensure the reproducibility and complete transparency of scientometric analyses, it is imperative that citation data is openly available (Sugimoto et al., 2017). However, the leading sources of citation data, such as the Web of Science (WoS) database and Scopus, still remain proprietary, limiting data access. A large number of researchers are at a disadvantage if they or the institution they are affiliated with cannot afford subscriptions to these data sources. Hutchins (2021) notes that, in the early years of citation indexing, maintaining the index was an arduous task and required substantial outlier funds. It was not feasible to provide free access to the data source due to insufficient public investment. Of late, data indexing is no longer a laborious process, and the cost involved has come down considerably.

There is a general acceptance that the papers cited by many others have had a more significant impact on other researchers than those with fewer citations. As Belter (2015) points out, the fundamental concept behind all bibliometric indicators or citation-based metrics is that we can assess the impact of a paper by quantifying the number of other papers that have referenced it. Lisciandra (2022) defines citation-based metrics as “statistical measures of scientific outputs that draw on citation indexes”. According to Mason and Singh (2022), citation-based metrics are frequently employed as a proxy for the prestige, quality and impact of academic journals, papers published in these journals, and the scholars who produce those papers. Though citation metrics are traditionally used in natural sciences, Lisciandra (2022) mentions that they are becoming increasingly relevant in the humanities, as well. However, there is a lot of controversy surrounding the bibliometric indicators and a debate over what they actually measure (D. W. Aksnes et al., 2019; Belter, 2015; Besselaar & Sandström, 2019).

Joshi (2014) classifies bibliometric indicators into three types, namely, quantitative, performance and structural indicators. The quality of journals and researchers is measured by performance indicators. These indicators are often used to determine rankings for entities like journals (Pajić, 2015). Various methods for ranking have been proposed, but there is no universally accepted method. However, the most popular method is considered to be the Impact Factor (IF) (Ritzberger, 2008). Many critics (e.g. McKiernan et al., 2019; Setti, 2013; Varin et al., 2016) have pointed out the limitations of using IFs as a tool for research evaluation. This metric will be discussed in detail in the coming chapters.

1.1. Motivation

Given the growing number of academics engaged in research and the substantial volume of academic papers published worldwide, citation analysis has become increasingly challenging. Citation analysis may appear straightforward. However, it involves several prerequisites and challenges.

As we will demonstrate in later sections, selecting the right citation database is crucial for citation analysis. Technological advancements have led to the development of several citation databases in recent years. It is imperative to note that each database has its own characteristics regarding coverage, the accuracy of the data indexed, the subjects covered and the timeliness of the data.

As part of the 'Initiative for Open Citations', scholarly publishers are urged to release the cited references in their journal articles to the public without any restrictions (Sugimoto et al., 2017). This initiative has promoted the development of a number of so-called 'open' sources of citation data. Also, due to the affordability and accessibility of open citation data sources, a recent shift towards these databases has been observed. Even though there are a number of studies involving older open sources of citation data, such as Crossref, studies on recently developed ones, such as Semantic Scholar and OpenAlex, are relatively few in number (e.g. Arum, 2016; Fricke, 2018; Kamińska, 2017; Scheidsteger & Haunschild, 2022).

The process of gathering relevant citation data is an important step in citation analysis. The data can be extracted from citation databases in a variety of ways, as we will see later. An efficient method to extract large amounts of data is to use Application Programming Interfaces (APIs). Organisations maintaining the data sources typically provide Representational State Transfer (REST) APIs to support data extraction. As each database has its own format for creating queries, different request parameters, and response formats, large-scale extraction of citation data from multiple databases is time-consuming and challenging without automation. Moreover, obtaining data using APIs requires familiarity with the data source's documentation on API usage. Several tools that support automated citation data retrieval (mostly developed by third parties) are available for use. But, these tools are mostly database specific and have different interfaces. Apart from having to install multiple tools, a significant drawback is that the user also needs to acquaint themselves with all of the database-specific tools they intend to use.

As we shall see in Chapter 4, most of the existing tools do not support the aggregation of citation counts at an entity level i.e., produce comprehensive citation reports for a set of entities like journals or authors for a specific time period. In these cases, a cross-citation matrix or table of cross-citations can be a useful format for the aggregation of citation counts and the generation of detailed reports. There is a cross-citation relationship between entity A and entity B if one of A's papers is cited by one of B's. A cross-citation matrix can be used to analyse and visualise the number of citations between entities. The matrices are a useful tool for measuring and characterising the information flow among entities (such as journals or authors) at various levels of aggregation (L. Zhang & Glänzel, 2004). The cross-citation matrices have been in use for the past several years for deriving citation-based metrics for ranking journals. As observed by Todorov and Glänzel (1988), "the tabulation of journal references to other journal titles (or the recording of the article-to-article links) in a cross-citing matrix is used (implicitly and explicitly) for deriving appropriate measures for journal interactivity and impact within an area of science and for specified periods of time". The author mentions that the matrix consists of rows and columns containing entity names (for example, journal titles), and each cell contains information about how many times a particular entity has been cited by another entity. The process of generation of a journal-to-journal cross-citation matrix involves, for example, searching for the required journals, retrieving the works in the selected journals, identifying the works that cite the works in the specified set of journals and finally, the aggregation of citation counts at a journal level. There is no denying that all of these processes are time-consuming and resource-intensive. Performing these procedures manually, that is, without automation, has a high probability of producing incorrect results. As a result, the development of a new tool that can overcome the challenges associated with the extraction and aggregation of citation data is essential in order to simplify, enhance, and improve the process.

1.2. Aim

The number of studies focusing on the open citation data sources developed recently is relatively small (e.g. Arum, 2016; Fricke, 2018; Kamińska, 2017; Scheidsteger & Haunschild, 2022). We aim to explore the use of newly developed open data sources to determine whether they compare with traditional/established sources of data such as Scopus. There may be different definitions of the term 'open data sources' among different individuals. For the purpose of this dissertation, 'open data sources' refer to those sources that provide citation data for free to the public through web services or in bulk, often under permissive licenses (Hutchins, 2021). While there are various licenses available, the most liberal option is a public domain license, which allows for unrestricted use and distribution of the data without any copyright restrictions. Commercial use may be restricted depending on the license, as per the author.

We developed 'CitaTrack', a Python-based package, to fill the void of a tool capable of extracting citation information across multiple data sources. The package also provides a uniform interface for metadata retrieval, thereby addressing the lack of a common interface for retrieving data from multiple sources. The primary purpose of the tool is to automate the extraction of citation data and to aggregate citation counts between different entities, such as journals or authors. Five different citation databases are used as sources of data for the development of the tool, namely Scopus, Crossref, OpenCitations Corpus, OpenAlex, and Semantic Scholar. The selection of sources was made considering their multidisciplinary nature, which allowed for the retrieval of citation data from a broad spectrum of research areas and subjects.

Citation analysis is a valuable research tool that can provide insights into the influence of scholarly works, but it comes with several challenges that must be addressed. A few of these challenges include issues related to the coverage of data, its availability, and the changing environment of research activities. Therefore, CitaTrack was developed as a modest attempt to address some of the challenges. The issue of varying coverage across databases is addressed to some extent if we use data from multiple data sources. Availability of data is yet another issue, as many of the widely used databases for citation analysis are proprietary. To promote the use of open citation data sources, four out of five sources of data used for the development of the tool are open. Moreover, the tool has been designed to accommodate the inclusion of additional data sources to address rapid changes in the research environment.

Additionally, citation analysis and ranking of journals were performed using the data obtained from the tool to illustrate the potential application of the tool or one like it in the field of bibliometrics. The main objective of this particular example was to rank the top 20 economic journals based on the cross-citation data obtained via CitaTrack using the Stigler model and to examine whether the resulting rankings agree with the general consensus about the best journals in economics. In addition, the resulting rankings are compared with other citation-based metrics, such as the IF and Article Influence Score. We have also highlighted the benefits of using multiple databases and citation-based metrics for research evaluation.

1.3. Structure

The thesis is divided into nine chapters. A brief introduction to bibliometrics and citation analysis is presented in the first chapter. The second chapter of the thesis gives the reader an overview of the topic and provides more insight into the field of bibliometrics, its origin, importance and citation analysis. Also discussed in detail are the criticisms and debates surrounding the use of citation-based metrics to measure performance and impact.

As we will see in the following chapters, selecting a data source prior to conducting citation analysis is a critical step. The third chapter provides an overview of the different sources of citation data and methods of obtaining citation data from these sources. A review of previous studies concerning the comparison of different databases is also presented. In these studies, the main characteristics and benefits/disadvantages of each database are discussed. This information is crucial for choosing the appropriate database for conducting citation analysis.

A discussion of existing tools for citation data extraction is presented in Chapter 4. These tools, however, have some limitations. In response, we developed CitaTrack to address these limitations. CitaTrack is a tool for extracting citation data and aggregating citation counts between a set of journals/authors. It offers several advantages over the pre-existing tools, and these advantages are explained in detail in Chapter 4.

Performing citation analyses is a common practice nowadays to determine which researchers or journals are the most influential in a given field. An objective measure of the impact of scholarly works is provided by citation-based metrics. The values of these metrics can then be used to rank different journals and authors. A discussion of various citation-based metrics and one of the statistical models, namely, the Stigler model, is presented in Chapter 5. Additionally, we explain why the Stigler model is superior to other citation-based metrics.

CitaTrack, the tool developed in this dissertation, produces journal-to-journal or author-to-author cross-citation tables as the output. Cross-citation data is a valuable source of information for researchers, providing insights into the relationships between different journals and authors within a particular field. The data can be used for a number of purposes, e.g. to rank a set of journals or authors or to perform a citation analysis. In chapters 6 and 7, a detailed discussion of an example of citation analysis and journal ranking based on the cross-citation data of 20 economic journals obtained with the tool is presented. Chapter 6 talks about the citation analysis & distribution of works in the set of journals, while Chapter 7 talks about the modelling of the cross-citation data using the Stigler model. The output obtained from the tool, that is, the journal-journal cross-citation data was fed into the Stigler model, and a set of 20 economic journals was ranked based on the data obtained from different citation databases.

Chapter 8 outlines the architecture of CitaTrack, its implementation and its data sources. In addition, the tool's usage and best practices considered during its development are briefly described.

The last chapter talks about the limitations of the study, the implications of this research for the academic community and concludes the dissertation.

2. Background

The emergence of bibliometrics as a field of study can be traced back to the 1920s and has significantly advanced since the 1960s (Cox et al., 2019). An overview of the field of bibliometrics & citation analysis, its origins and a brief summary of open access journals are presented in this chapter.

2.1. Bibliometrics

The term ‘bibliometrics’ was coined by Pritchard (1969) and the author describes it as “the application of mathematical and statistical methods to books and other media of communication”. Back in the early twentieth century, librarians used bibliometrics and citation analysis for the management of journal collections (Gingras, 2016). The bibliometric references were used to determine the usefulness of the papers and journals. Further, the rapid growth of research literature led to the development of new methods of structuring it. A pioneer in the field of bibliometrics, Eugene Garfield, introduced the idea of citation indexing (Prathap, 2017). The development of the Science Citation Index (SCI) by him revolutionised the way scientific literature was retrieved & tracked, and the process of counting and tracking citations to individual articles was also systematised (Thompson & Walker, 2015).

Zupic and Čater (2015) observe that bibliometric methods are used primarily for science mapping as well as performance analysis. The main objective of science mapping is to uncover the structure and dynamics of scientific fields, whereas performance analysis is concerned with evaluating research, institutions, and researchers. Nowadays, a scientific work’s impact or influence is commonly estimated using bibliometric indicators/metrics and is based on the number of citations it receives (Cooper, 2015). The most commonly used citation-based metrics will be discussed in Chapter 5.

2.2. Citation Analysis

There has been an increasing interest in the field of bibliometrics regarding citation analysis over the past few years (Qiu et al., 2017). Citation analysis focuses on the relationship between a paper and its references and is based on the assumption that these articles might be related (McBurney & Novak, 2002).

Scientific tradition requires that when a reputable scientist or technologist documents their research, they should refer to earlier articles which relate to the subject of the documented work (Nicolaisen, 2007). According to the author, the scientist or technologist may have drawn inspiration from or used the concepts, theories, and methods of earlier researchers while conducting and presenting his or her research and the bibliographic references in work make it possible to identify the influential researchers. Therefore, researchers often use citations as a means to pay homage to pioneers, identify original publications, correct the work of others, alert readers to forthcoming works, verify data, provide background information or reading and give credit to related works (Garfield, 1965). Also, citations bring in more transparency and reliability to research work and are a useful tool to avoid plagiarism. It also provides them with an opportunity to have a deeper look into the subject under discussion.

The main elements of the bibliometric analysis as observed by Thompson and Walker (2015) are coverage of the database, data fields, search options, consistency & accuracy of the data, and analysis and the usage of metrics. Database coverage is an essential factor to be considered while choosing the data source for citation analysis as each database may have different coverages relating to subject areas covered, languages, etc. The accuracy of citation data is another aspect to be considered. Many factors contribute to citation database errors, including mistakes committed by the authors when listing their references and errors made during the data entry process (Buchanan, 2006). Appropriate database relevant to the subject of citation databases needs to be carefully selected as data fields may differ from one citation database to another. The different search options offered by each database should be assessed beforehand to choose the most relevant database for the citation study. Yet another element to be considered is the analysis and usage of citation-based metrics. As Lisciandra (2022) notes, "citation metrics are statistical measures that combine citation data with other variables, for instance, citations over periods of time or citations over quantity of publications". A detailed discussion of the citation-based metrics is provided in Chapter 5.

Todorov and Glänzel (1988) note that citation-based metrics can be derived from cross-citation matrices. According to the author, to obtain these metrics, one can either include or exclude the number of papers published by the journal (i.e., the size of the journal) and vary the source journals, the cited journals, and the time period of the matrix. Also, cross-citations between journals/authors are an important component of scientific research since it shows how one author's/journal's work influences or impacts another's. A cross-citation matrix is a key tool for exploring the relationship between the set of entities and their works. It provides a visual representation of the connections between entities and their works, allowing us to better understand how their works interact with each other. For instance, an author-to-author matrix can be used to identify clusters of authors who cite each other frequently. Table 2.1 is an example of an author-to-author cross-citation matrix. An analysis of the cross-citation matrix can also reveal which journals/authors are being cited most frequently and which have the greatest influence. Moreover, these matrices are useful for tracking trends over time (Sainaghi et al., 2018).

Table 2.1.: *An author-to-author cross-citation matrix for articles published in 2015-2023*

	Bastian Greshake Tzovaras	Daniel Himmelstein	Jacob G. Levernier
Bastian Greshake Tzovaras	51	2	1
Daniel Himmelstein	8	146	2
Jacob G. Levernier	4	2	2

Over the past few decades, numerous citation indicators have been developed, leading to an extensive debate surrounding the appropriate methods for calculating these indicators, the coverage of the database used, the normalization procedures used and the quality of the underlying data (D. W. Aksnes et al., 2019). Ever since these indicators were used to evaluate scientific performance, it has been the subject of significant discussion and there have been controversial opinions regarding the acceptability of these metrics as performance or impact indicators. Lisciandra (2022) suggests that rankings based on citation metrics, which can range from individual scientists and articles to entire universities, rely on the idea that citations reflect specific positive aspects of scientific research.

However, MacRoberts and MacRoberts (1989) draw attention to inherent problems with citations like variation in citation rates, citations pointing out incorrect results, citation practices that are biased, errors in citation data sources and self-citations. Authors like Jensenius et al. (2018) believe that citation practices put certain scholars at a disadvantage, including those with innovative ideas, early-career academics, researchers working in smaller communities, women, and solo authors. The citation count, therefore, in their opinion, may exacerbate social hierarchies and inequalities by providing an inaccurate assessment of the impact. On the other hand, there is a general consensus within bibliometrics that citations are a good but not the perfect measure of impact (D. W. Aksnes et al., 2019).

Even though a part of the scientific community views the measures for evaluating journals or works derived from citation analysis with scepticism, it is still considered an important and valuable tool to judge the quality and impact of research work. Over the past three decades, citation analysis has gained popularity as a means of evaluating scientists and their research, in combination with peer review practices (Meho, 2007). Many bibliometricians believe that citation analysis or citation-based metrics should not be considered as a replacement for peer review due to the various limitations of the citation analysis (D. W. Aksnes et al., 2019). Peer review methods have their own weaknesses since this can be a time-consuming and costly process, and it often fails to produce consistent or reproducible results (Belter, 2015). Accordingly, the better alternative will be the complementary use of citation analysis methods along with peer review methods.

2.3. Open Access Journals

The term ‘Open Access’ (OA) refers to the free online access of peer-reviewed research journal articles (Harnad, 2015). Researchers are subject to many restrictions when it comes to accessing scientific articles. A major issue is the payment of subscriptions, which results in a financial burden to the individual researcher unless he or she belongs to a prestigious institution that provides access to a large number of journals. In addition, no institution can afford the subscription fees for all journals. In the modern era, where large amounts of data are available at our fingertips free of charge, the fact that a large number of publications and articles are unavailable to researchers without subscriptions to various journals is disappointing.

These issues have led to the emergence of the OA movement, and it is gaining momentum in the research community (Piwowar et al., 2018; Suber, 2012; Zhixiong et al., 2013). A key objective of the OA movement is free and unrestricted access to primary research literature. OA has many advantages for research, including increased visibility, fostering innovation in the industry, and reducing financial constraints on academic and research libraries (Tennant et al., 2016). Moreover, researchers and non-academics will have equal access to research output and studies. According to the studies conducted by Piwowar et al. (2018) and Eysenbach (2006), OA articles, in general, receive more citations than non-OA articles.

OA models can be distinguished based on the type of access to scholarly materials provided to the reader. Articles published in Gold OA journals are completely free for readers (Harnad, 2015). However, authors or institutions must pay an article processing charge for the publication of articles in the Gold OA journals. The author notes that investing in Gold OA can further increase the financial strain on institutions that are already struggling to bear the subscription costs of journals. In green OA, articles are self-archived in openly accessible repositories. A hybrid OA journal combines OA and subscription-based models, which means that the journal publishes non-OA articles but offers authors the option to pay Gold OA fees and make their articles freely available. Piwowar et al. (2018) observe that OA articles are more likely to be recognized and referenced by peers compared to non-OA articles that are published in the same journal. Diamond OA, on the other hand, does not charge authors for publishing and is funded by alternative sources.

As Suber (2012) observes, OA benefits researchers and non-researchers alike. It helps in the facilitation of research and the widespread dissemination of results. OA supports readers to easily find and retrieve the information they need and, on the other hand, helps the authors to reach out to readers who can utilize, reference, and build upon their work. The advocates of OA journals (e.g. Harnad, 2015; Piwowar et al., 2018; Suber, 2012) are of the opinion that the articles have increased visibility, wider reach among the audience and have a citation advantage compared to non OA journals.

On the other hand, many scholars (e.g. Beall, 2013; Xia et al., 2014) believe that the editorial standards of OA journals are often lower than those of traditional journals, and unscrupulous authors and publishers are more likely to take advantage of them. They observed that many OA journals are of inferior quality and lack transparency in terms of peer review and publication fees. These journals are commonly referred to as 'predatory' journals. Bohannon (2013) submitted 304 variations of one of his papers to OA journals as part of his study. However, more than half of the journals accepted his paper without noticing most of the flaws in his paper. According to the author, a significant number of these journals do not conduct proper peer reviews.

Overall, there are differing opinions on the specifics of OA implementation and a general consensus on the topic is yet to be reached. As the debate continues, it is important to consider both the potential benefits and drawbacks of open access.

3. Sources of Citation Data

Citation databases are valuable resources for academics, as they provide access to a vast collection of reliable and peer-reviewed research materials. In this chapter, different sources of citation data will be discussed, as well as the pros and cons of using each database. Additionally, we examine the different methods of extracting data from them.

Citation databases contain information regarding scholarly materials and their citations stored in a structured and consistent way. Citation data includes information such as journal name, date published, DOI, author, citation counts etc. One of the first citation databases (WoS) was developed by ISI (now Clarivate Analytics) during the 1950s. It was in the early 2000s that the Crossref & Scopus databases were launched. For a long time, the market was dominated by Elsevier's Scopus and the WoS database. New citation databases like OpenAlex, Semantic Scholar, Microsoft Academic etc., were developed in recent years. However, access to proprietary databases like Scopus and WoS is restricted, while the other databases can be accessed free of cost.

3.1. Extraction of Data

The three most common ways to extract metadata from a citation database are the following:

- REST APIs: The infrastructure, specifically the API endpoints to retrieve data, is already set up by the organisation maintaining the citation databases. The APIs can then be used to retrieve the required metadata. All the citation databases used for developing CitaTrack, the tool developed, have the provision to retrieve metadata using APIs.

- Database Snapshot: Some organisations allow us to download the snapshot or copy of the entire database. The snapshots are updated regularly, usually biweekly or monthly. However, additional tools are required for the large-scale mining of the data.

- Website: Almost all organisations have a web interface where users can retrieve the required data using a simple search. These web interfaces can be used for data extraction and citation analysis on a relatively small scale. It is, however, necessary to use either APIs or database snapshots for data extraction on a larger scale.

3.2. Study of Citation Databases

Database coverage heavily influences the citation data that can be retrieved since the database can only include citations from items indexed by it (Bar-Ilan et al., 2007). As these databases may have different coverages, there is an increasing need to systematically compare the different citation databases. Studies comparing the citation databases are discussed in detail in section 3.3.

Using CitaTrack, the tool developed in this dissertation, citation data of the top 20 journals in Economics were retrieved from 5 different citation databases, and the journals were then ranked based on the data obtained. The procedure is discussed in more detail in Chapter 7. The following citation databases were analysed in the initial phase to obtain the data for citation analysis.

3.2.1. Web of Science Database

WoS database is among the oldest and most widely used citation databases within the field of bibliometrics. The idea of citation indexing for the sciences was originally proposed by Dr Eugene Garfield, the founder of ISI (now Clarivate Analytics) in 1955, and the first Science Citation Index (SCI) was created by ISI in 1964 (Clarivate, 2022). Other indexes of WoS include Social Sciences Citation Index, Arts & Humanities Citation Index, Conference Proceedings Citation Index, Book Citation Index and Emerging Sources Citation Index (Singh et al., 2021). According to Clarivate (2022), the Journal Citation Report was launched to collect information on citations between academic journals and this was done to aid publishers and librarians in gaining a better understanding of the scientific and social sciences literature communication network, as well as the reputation and impact of particular journals. With more than 5,600 journals spanning more than 150 scientific disciplines (Clarivate, 2023), Clarivate Analytics (formerly Thomson Reuters; ISI) has established itself as one of the domain leaders.

According to a study by Birkle et al. (2020), access to WoS data for institutions and affiliated researchers is facilitated through various channels such as platforms, APIs, and custom data set delivery. One can update and re-extract these data sets once a year or as frequently as every two weeks, depending on their needs. Furthermore, the authors observed that WoS provides three types of API formats (basic, intermediate, and advanced formats) that can be tailored to meet the specific requirements and technical expertise of the customers in terms of call speeds and data volumes. The WoS data is the source of several commonly used citation-based metrics, such as the IF and Eigenfactor. Despite many criticisms, these metrics are still considered the gold standard in traditional bibliometrics (Kovatcheva, 2022).

In our correspondence with the WoS team, we were informed that a subscription to the database is required in order to utilise the APIs offered by them. Data extraction from the WoS database could not be supported in CitaTrack due to the lack of a subscription to the database by the RPTU.

3.2.2. Crossref

The Crossref database was founded in January 2000 by a group of publishers seeking an efficient means of linking journal articles (Hendricks et al., 2020). Digital Object Identifiers (DOIs) were used to link references between articles, making it easier for a reader to locate cited items (Collins, 2018). Persistent identifiers, specifically, DOIs, contain a resolution service and metadata about the resources being referred to and are primarily managed by Crossref through a registration process (Chudlarský & Dvořák, 2017). Using a DOI to identify an item provides a unique and permanent link that stays with the item even if its website changes (Collins, 2018). The organisation's current formal mission statement reads (Crossref, 2023):

"*Crossref makes research objects easy to find, cite, link, assess, and reuse. We are a not-for-profit membership organisation that exists to make scholarly communications better.*"

In 2006, the service 'cited by' was released, enabling the members to retrieve citation counts for their works and the APIs [1] were released in 2013, making their metadata public and license-free (Hendricks et al., 2020). In order to extract metadata, either their website can be used to look up a small number of DOIs, or APIs can be used to search and retrieve metadata for a list of DOIs. It is possible to download a database snapshot as well. However, it is only available to metadata Plus users. Metadata Plus is a paid service offered by Crossref, which gives the users enhanced access to the Crossref APIs, improved service and additional features such as priority service/rate limits.

Today, Crossref is considered an important source for retrieving metadata in the field of bibliometrics. With a growth rate averaging 11% per year, Crossref's database now boasts more than 106 million records. As a result, the metadata provided by Crossref has emerged as a significant source of scholarly data (Hendricks et al., 2020). Also, the number of open citations in the database is on the increase, and more than 50% of the citations in the Crossref database were classified as open (Shotton, 2017). Meanwhile, a study conducted by Chudlarský and Dvořák (2017) suggests otherwise. It was found that the coverage of the Crossref database was lesser than that of the WoS databases. Contrary to the popular opinion they argued that the extent of citation coverage offered by Crossref is insufficient to rely on it as the primary source for the analyses of citations in research evaluations conducted at the university and faculty levels.

[1] https://api.Crossref.org/swagger-ui/index.html

3.2.3. Elsevier's Scopus database

Elsevier launched Scopus in 2004 as a peer-reviewed abstract and citation database that has continuously been updated since then. It began with around 27 million records and grew to 76 million by 2019, making it one of the largest abstract and citation databases (Baas et al., 2020). Currently, only rigorously selected publications are indexed in the database, which means the content is highly curated.

A study by Singh et al. (2021) highlights that the Scopus platform provides the option of searching, discovering, and analysing data. Various search options include document, author, and advanced searches. Users can utilise the Discover feature on the Scopus platform to locate related publications using metrics like author keywords & shared references and identify collaborators and research organisations based on their research output. Moreover, the authors state that the citations can be tracked, and search results can be analysed based on criteria such as country-wise, affiliation-wise, and research area-wise distribution using the analyse option.

Scopus is a multidisciplinary database, and one can easily search for documents, authors or affiliations on their website. In order to programmatically access data, an API key is required. Elsevier's developer portal allows one to generate an API key, which should be passed as part of the request header while sending a request for data retrieval[2]. Elsevier's policies regarding the use of APIs and data can be found on their website. For non-commercial use, everyone can obtain API keys free of charge, subject to Elsevier's API and Data usage policies. However, anyone using the APIs for commercial purposes will need a dedicated API subscription (B.V., 2022). However, non-subscribers can only access limited metadata from the Scopus database and has access only to basic search functionalities. Advanced search and information such as citation count is provided only to subscribers of the database.

3.2.4. Microsoft Academic Graph

The Microsoft Academic graph was released in the year 2015. Sinha et al. (2015) describe it as "a heterogeneous entity graph comprised of six types of entities that model the scholarly activities: field of study, author, institution, paper, venue, and event". One could either download the database snapshot or use the APIs provided by the organisation to extract the metadata. A number of technologies were used to enhance the exploration of academic information, including data mining, machine learning, and semantic analysis (Wan et al., 2018).

The study by Herrmannova and Knoth (2016) found that the MAG has ex-

[2] https://dev.elsevier.com/apikey/create

cellent coverage across different domains. The most prominent fields of study were computer science & engineering, chemistry, physics and biology. As a result, it was said to have a bias towards the technical disciplines. It could be attributed to the reason that MAG uses web crawlers, and comparatively, more works from the scientific disciplines are published online. In the study, certain limitations were identified, including incomplete data, a lack of availability of affiliation information, and the absence of normalisation for institution names.

The database had an extensive collection of metadata and comprehensive coverage, making it a promising database (A. W. Harzing & Alakangas, 2017). Herrmannova and Knoth (2016) argue that, with more than 120 million publications and related authors, venues, organisations, and fields of study, the Microsoft Academic Graph was the largest publicly available dataset of scholarly works and the largest dataset of open citation data back in 2016. In their opinion, "Microsoft Academic graph was the most comprehensive publicly available dataset of its kind and represents an astonishing effort which will prove useful in many areas of research where full-text access to publications is not required".

However, in May 2021, the retirement of the Microsoft Academic website, as well as its application programming interfaces and snapshots, by December 31, 2021, was announced via Microsoft Blog (Scheidsteger & Haunschild, 2022). It was considered difficult to replace Microsoft Academic Graph with an alternative. It received updates twice a week until the end of 2021.

Post the announcement of the retirement of MAG, a non-profit research organisation called 'OurResearch' announced that they would preserve and incorporate the most recent complete MAG corpus, with the exception of patent data, while also striving to enhance it (Scheidsteger & Haunschild, 2022). Their motto was to provide "a fully open catalog of the global research system" (OurResearch, 2021). Shortly after the MAG's retirement in December, a citation database called OpenAlex was launched in January 2022.

3.2.5. OpenAlex

OpenAlex is a relatively new database that was launched in 2022. The inspiration behind the name 'OpenAlex' came from the ancient library of Alexandria in Egypt, which was renowned for maintaining the world's first catalog of library collections (Piwowar et al., 2022). As mentioned in section 3.2.4, it was launched as a replacement for the Microsoft Academic Graph post its retirement. Replacing one of the largest open databases was considered difficult, and OpenAlex was developed to address this concern (Priem et al., 2022).

It mainly contains five types of entities - sources (venues), concepts, works, authors and institutions, which are connected, forming a heterogeneous graph. Figure 3.1 shows the OpenAlex graph illustrating the relationship between entities. All entities have an OpenAlex ID - a unique, non-nullable identifier that acts as the database's primary key. The ID is in the form of a URL, which is both human and machine-readable. There are currently three ways to retrieve the data from the database:

1. Downloading the entire data dump: The data dump is updated fortnightly
2. Using REST API calls: The metadata is updated daily and has no rate limit.
3. Using a web-based GUI which is built on the REST API.

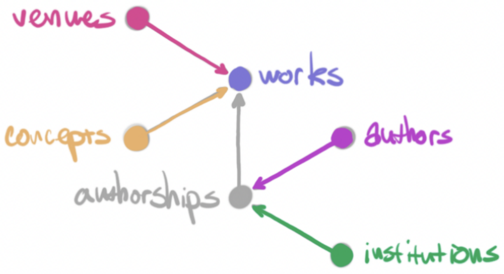


Figure 3.1.: *Sketch of the OpenAlex graph data model (Priem et al., 2022)*

The primary sources of the database are Microsoft Academic Graph and Crossref. In a study conducted by Scheidsteger and Haunschild (2022), it was found that almost all works from MAG have been transferred by OpenAlex while maintaining their bibliographic data, first and last page, volume, publication year, DOI, and the number of references which are crucial for citation analysis. The document types in OpenAlex were 90% similar to that of MAG. Also, OpenAlex has introduced an additional document type, namely journal article, which attributes to more than 7% of all document type specifications. The authors point out that, overall, the OpenAlex database appeared to be better suited for bibliometric analyses compared to MAG.

Priem et al. (2022) remark that OpenAlex has the potential to improve the transparency of research, evaluation, navigation, representation, and discovery despite being a fully open (100% open data, open API, open-source code) source of scholarly metadata.

However, it is a relatively young database. Therefore it has some limitations. The major contributors to the development of the database, Priem et al. (2022), acknowledged the limitations of the OpenAlex database, specifically regarding the normalisation, parsing and disambiguation of entities such as authors and institutions. In addition, there is a lack of metadata on funding sources and corresponding authors in the database. The authors also point out that the accuracy and completeness of the database are yet to be determined. Despite its limitations, it remains a promising database.

3.2.6. OpenCitations Corpus

OpenCitations was founded in 2010 at the University of Oxford with the aim of providing open bibliographic citation information in the Resource Description Framework (RDF). RDF is used to describe resources and their relationships, which can be stored in different sources, and linked together to form a network of interconnected information called the Semantic Web (Peroni & Shotton, 2020). It was initially started as a one-year OpenCitations project, and later in 2015, following a formal agreement by the University of Bologna and the University of Oxford, a newly created version of the OpenCitations Corpus (OCC) based on a revised metadata schema was built from an initial prototype by the University of Oxford (Peroni et al., 2017). OpenCitations is an organisation focused on promoting open scholarship and making open bibliographic and citation data available using Semantic Web technologies (Heibi et al., 2019). The organisation has developed a scholarly infrastructure to support these goals and continues to work towards greater openness and transparency in scholarly research. Open data is beneficial, especially in the research field, as other researchers can quickly reproduce results from other previous research and investigate them further.

The COCI project is a valuable open alternative to commercial citation databases like WoS (by Clarivate Analytics) and provides openly accessible citation links from Crossref, which are marked as open and free to use (Chudlarský & Dvořák, 2017). Unlike traditional databases, the access to which is often limited by paid licenses, the COCI project offers an open source of citation data that is available free of cost to researchers worldwide.

Peroni et al. (2017) note that the OCC stands out as the most extensive and ‘truly open’ compilation of citation data in RDF format accessible on the internet. In this context, ’truly open’ refers to all citation data being freely and publicly available, since publishers can make reference lists private in some databases, such as Crossref. As noted by Peroni and Shotton (2020), the methods using which the metadata in the OCC database can be extracted are:

1. The metadata can be accessed through SPARQL endpoints using appropriate SPARQL queries.

2. One can download the data in JSON and CSV formats via the REST APIs, which has been developed using RAMOSE (Restful API Manager Over SPARQL Endpoints).

3. The data dumps, which are updated on a monthly basis, are maintained in JSON-LD format and stored online with Figshare’s assistance.

4. The bibliographic entities can be obtained in multiple file formats through content negotiation by utilizing the HTTP URI of each individual entity.

5. To locate and browse bibliographic entities, OpenCitations has devised OSCAR and LUCINDA, interfaces for search and browsing.

3.2.7. Google Scholar

Google Scholar was released in 2004. As a well-known academic search engine, Google Scholar provides access to a vast array of scholarly literature from various disciplines and databases (Arum, 2016). It works in a similar way to Google's main search engine, by producing search results that are determined by the strength of the link between the search terms and how frequently and recently a particular work has been referenced (Jensenius et al., 2018). Google Scholar provides a simple and easy-to-use platform to explore the scholarly literature, allowing users to browse related works, citations, publications, and authors, locate the full document online, and stay up-to-date with the latest developments in any research field (Scholar, 2022).

Various advantages of the search engine were noted in a study conducted by Jensenius et al. (2018). The main advantage of Google Scholar is that it is easy to use. By providing a comprehensive list of a scholar's publications ranked by their number of citations, Google Scholar offers a convenient way to get an overview of their work. Users can also click on the hyperlinks of each publication to view abstracts and gain access to articles that are publicly available. As observed by the authors, Google Scholar has several benefits for research evaluation, including promoting consistency in evaluation practices, encouraging transparency, publicity, and openness in scholarly work. It also facilitates access to scholarly materials, helps to connect scholars and encourages networking among them. Additionally, Google Scholar may provide incentives for quality over quantity in research output. On the other hand, the authors remark that google scholar's citation counts tend to favour scholars who produce incremental research, work in larger research communities, are male (and likely white), collaborate with others on their research, and receive strategic citations.
Other advantages as noted by Zientek et al. (2018) are:

1. One can easily track an academic's research through their Google Scholar Profile.

2. It can facilitate the identification of a set of articles that focus on a particular subject.

3. It provides historical trends in research, that is, one can easily track research over time for a publication or a scholar.

4. It promotes meta-analytic studies.

5. It helps to bridge the distance between scholarly research & social media.

Waltman (2016) points out various issues of Google Scholar. The search engine suffers from a lack of quality control. It is observed that there are many reports of inaccuracies in Google Scholar in the literature, e.g., problems related to content gaps, incorrect citation counts, and phantom data. Cleaning the data can be very time-consuming. Moreover, it is not feasible to get access to the complete Google Scholar database. Zientek et al. (2018) remark that conducting extensive citation analyses using Google Scholar can be challenging due to the fact that the database can only be accessed through its web interface.

3.2.8. AMiner

AMiner, an academic search and mining system, was launched in December 2015 and is the second generation of the Arnet Miner System (J. T. J. Zhang et al., 2008). Wan et al. (2018) mention that, the system's main aim is to provide a "systematic modelling approach to help researchers and scientists gain a deeper understanding of the large and heterogeneous networks formed by authors, papers, conferences, journals and organisations". The profiles of different researchers are extracted automatically from the web and are integrated with the papers published by them after a name disambiguation is performed (Tang, 2016). The author also mentions that AMiner offers a set of researcher-centred functions, including collaboration recommendation, social influence analysis, similarity analysis, influence visualisation, community evolution, and relationship mining.

According to Wan et al. (2018), AMiner focuses on:

1. extracting information from the distributed web to develop a semantic-based profile for each researcher.
2. combining data such as researcher profiles and bibliographic data, the system integrates data from multiple sources.
3. performing precise searches within a heterogeneous network.
4. conducting an analysis to uncover interesting patterns within the constructed social network of researchers.

The author also remarks that there is currently no consistent process or set of methods for extracting data from various academic social networks. Therefore, AMiner was developed to conduct data mining and search operations on academic publications available on the internet, utilizing social network analysis to recognize the relationships between researchers, publications, and conferences. The citation data are extracted from data sources including Association for Computing Machinery (ACM) digital library, and Digital Bibliography & Library Project (DBLP).

Several different datasets can be downloaded on the AMiner website [3] for free.

[3] https://www.aminer.org/data/

Downloading and mining their entire database would need additional resources and tools. Also, there is a fee associated with the usage of their API endpoints for fetching the metadata for published works. Hence, the citation database could not be integrated into CitaTrack.

3.2.9. Semantic Scholar

Semantic Scholar was developed in 2015 by the Allen Institute for Artificial Intelligence. Their main aim is to help scholars combat information overload and more efficiently discover and understand the most relevant research literature. It was designed to be an intelligent search engine, assisting researchers in finding high-quality academic publications more quickly and efficiently (Wan et al., 2018).

Wade (2022) provides an overview of the Semantic Scholar database in his paper. The author describes the Semantic Scholar Academic Graph, as a “large, open, heterogeneous knowledge graph of scholarly works, authors, and citations that powers the Semantic Scholar discovery service”. With more than 205 million publications, 121 million authors, and nearly 2.5 billion citation edges, it is an extensive and comprehensive repository of scholarly information. It integrates metadata from various sources, such as PubMed, Crossref, and Unpaywall, among others, providing a vast array of high-quality data to support scholarly research.

The metadata can be retrieved with the help of their APIs [4] provided by the organisation and via the downloadable database snapshots [5]. Nevertheless, the bulk download of data is only available for authenticated users. One has to fill out a form on their website to obtain the API key for authentication. Semantic Scholar employs machine learning techniques to analyse millions of research papers and helps researchers find better academic publications faster (Arum, 2016). They use natural language processing techniques to analyse publications, extract details and find the most relevant results, adding a supplementary semantic analysis layer to the traditional citation analysis methods. The advantages and disadvantages of Semantic Scholar compared to Google Scholar are presented in the next section.

3.3. Previously published studies comparing Citation Databases

Traditionally, two proprietary databases: Scopus and WoS, were commonly used for citation analysis, and the competition between them has been intense. Recently, newer citation databases such as OpenAlex, Semantic Scholar, OCC,

[4] https://www.semanticscholar.org/product/api
[5] https://api.semanticscholar.org/datasets/v1/release/latest

and others have been developed that allow free and open access.

Technological development has helped in the creation and development of several citation databases. Some of these databases are multidisciplinary, like the WoS or Crossref database, while others are restricted to a particular domain, like the Pubmed Database. These databases have different coverages across different subject categories or languages, each with merits and demerits. Therefore, it is imperative that different factors such as subject & language coverage, reliability of the database, author disambiguation mechanism etc should be considered whenever a citation analysis is to be performed. The studies comparing the different databases help us determine the appropriate database for different use cases involving citation data.

Comprehensive studies covering all the citation databases mentioned previously are currently not available. Therefore, several studies comparing different subsets of the listed databases are given below:

Numerous studies have compared the WoS & Scopus databases extensively. One such study by Vera-Baceta et al. (2019) found that there is an over-representation of journals in English, in WoS, i.e., 95.37% and Scopus, i.e., 92.64%. In terms of the number of documents, Chinese had the second-highest representation with 2.76% in Scopus, whereas Spanish had 1.26% in WoS. It clearly shows the dominance of the documents in English over other languages, and the coverage of publications in other languages is relatively higher in Scopus compared to WoS, according to the authors. Many studies (e.g. D. Aksnes & Sivertsen, 2019; Mongeon & Paul-Hus, 2014; Vera-Baceta et al., 2019) highlight the bias of WoS and Scopus towards the English language and point out that both the databases have the lowest coverage in social sciences and humanities. Mongeon and Paul-Hus (2014) also remarked that the coverage of the Scopus was more than that of WoS.

An analysis of citation data from the Scopus, Google Scholar and WoS databases in 252 subject categories by Martín-Martín et al. (2018) revealed that Google Scholar found around 95% of WoS citations, 92% of the Scopus citations and a large number of unique citations. However, the unique citations found by Google Scholar had, on average, a much lower scientific impact than those found by WoS or Scopus. Additionally, it was noted that Google Scholar's citation data was essentially a superset of WoS and Scopus citation data, having a significant additional coverage for all areas. This might be attributed to the fact that Google Scholar automatically indexes publications and citations with the help of web crawlers, which results in significantly higher coverage compared to Scopus and WoS (Franceschini et al., 2016). However, the author believes that it is almost impossible for Google Scholar to compete with its two competitors because its automatic indexing causes many errors. On the other hand, the author mentions that some recent studies indicate that the data quality of Google Scholar is getting better over time.

Table 3.1.: *Comparison between Semantic Scholar and Google Scholar (Arum, 2016)*

	Semantic Scholar	Google Scholar
Advantages	Lower recall, higher precision	Higher recall, lower precision.
	Numerous filter options available: Field of study, publication year, author, key phrases, publication venue, the data set used.	Numerous filter options available: articles/case law, time of publication, patent, and citation inclusion.
	Sort capability for relevance or recency.	Sort capability for relevance or date.
	Links to references, citations, and related publications.	Result list multiple versions, if applicable.
	Extract core metadata from figures, tables and captions.	Advanced search available.
		Includes features such as 'My Library', 'My Citations', 'Alert', and 'Metric'.
Disadvantages	Fewer research results (limited to computer science and neuroscience).	Redundant search results.
	No recent citation information is listed in search results.	Sorting by date only applies to articles added in the past year.
	Lacks advanced search	Difficult to exclude articles that cite the author from the result list.

Another study by A.-W. Harzing (2019) compared Crossref, Dimensions (yet another open citation database), Scopus, Microsoft Academic and the WoS databases. The study's findings revealed that while Crossref and Dimensions provided similar or higher coverage for publications and citations than Scopus and WoS, their overall coverage was significantly lower compared to Google Scholar and Microsoft Academic. Additionally, it was observed that Crossref and Dimensions could be viable substitutes for Scopus and WoS for citation analysis and literature reviews. Consequently, the author also argues that Google Scholar and Microsoft Academic hold the top position as the most comprehensive and free sources for accessing publication and citation data. An almost identical result was obtained in the study conducted by Martín-Martín et al. (2016), and it was observed that Google Scholar had the upper hand over the other databases compared in the study. The percentage of citations found by Microsoft Academic was the second largest overall. However, it lacked coverage in subject areas like physics or humanities compared to the

WoS or Scopus databases. The Scopus database occupied the third position, and the COCI database had the smallest coverage regarding the percentage of citations in different subject areas.

Google Scholar and Semantic Scholar are academic search engines that can be used to search for scientific publications. A comparative study between Google Scholar and Semantic Scholar databases by Arum (2016) highlights the advantages and disadvantages of the two search engines. Table 3.1 highlights the main differences between both databases. Google Scholar has wide coverage across different subjects, whereas Semantic Scholar being a relatively young database, is limited to fields of Computer Science and Neuro Science. Moreover, it was noticed that the precision and relevance of Semantic Scholar's search results were higher than Google Scholar's. Also, the number of documents indexed in the former is much smaller than in the latter. Wan et al. (2018), have observed that Semantic Scholar is faster than Google Scholar and Microsoft Academic in highlighting the most significant papers and detecting the relationships between them.

Ambiguity in the name of different authors is a major issue to consider when conducting citation analysis. For example, different authors appear to have the same names and abbreviated forms of names make it difficult to link a publication to the original author. Many citation databases have author disambiguation systems, for instance, AMiner, Semantic Scholar and Google Scholar. A comparison of results of the author disambiguation mechanisms in AMiner, MAG and Semantic Scholar by L. Zhang et al. (2020) revealed that Semantic Scholar and MAG achieved a better performance than the AMiner. However, the study warrants the improvement of the author disambiguation mechanism in all three databases (L. Zhang et al., 2020).

Since OpenAlex is the most recent database, no studies comparing the OpenAlex database to other databases or studies discussing the coverage have taken place. The only exception is the study comparing metadata between Microsoft Academic Graph and OpenAlex (Scheidsteger & Haunschild, 2022), which has already been discussed in section 3.2.5.

4. Existing Tools for Citation Data Extraction

Manual retrieval of citation data can be a cumbersome and time-consuming procedure. Therefore, libraries or packages written in a programming language like Python, Ruby and R, are generally preferred and used for extracting data for citation analysis, making the extraction process simpler and more efficient than manual data extraction. Several libraries are available to extract metadata using APIs from different citation databases. Most libraries are user-friendly and can be used by someone with little to no knowledge about the usage of APIs. Having examined various citation data sources, let's take a closer look at some existing libraries and packages that can be used to extract metadata from citation databases. Python and R are the most commonly used languages for citation analysis; therefore, we will focus on libraries and packages developed in those languages.

4.1. Examples of the tools

For each citation data source used for the development of CitaTrack, an existing library or package that is widely used for data extraction or manipulation, is mentioned below.

4.1.1. Pybliometrics

Pybliometrics [1], a Python-based library, can be used to extract & cache metadata from the Scopus database and has a simple and consistent interface. The tool can be seamlessly integrated with various components in Python's data science ecosystem, such as machine learning and visualisation tools, without the need for a server (Rose & Kitchin, 2019). The tool can be downloaded from PyPI and can be used either within a Python interpreter or in script mode.

Scopus has 11 APIs in total, of which data retrieval using eight API endpoints is supported by the tool, i.e., the three retrieval APIs: Abstract, Affiliation & Author Retrieval, the three search APIs: Abstract, Author and Affiliation Search, and two metadata APIs: Citation Overview API & Serial Title API (Rose & Kitchin, 2019). When retrieving data, providing mandatory input parameters, such as search query strings (for search classes) or identifiers for

[1] https://pybliometrics.readthedocs.io/en/stable/

Scopus entities (for metadata and retrieval classes), is required. The Citation Overview API requires an additional parameter, i.e., the start date. Aside from the ease of use, the package offers the advantage of not having to browse the complete documentation of the Scopus APIs to extract metadata.

Metadata related to a set of works, authors, affiliations etc., can easily be retrieved via the tool, and to speed up future retrievals, the responses are also cached. The package supports error handling as well. Also, when used for the first time, the authentication details, i.e. the API key generated on Elsevier's website or the institution key, should be input. The authentication information is then saved in a configuration file for future use. Additional information related to the usage of the package can be found on the pybliometrics website[1].

4.1.2. openAlexR

One of the popular R-packages to gather citation data from the OpenAlex database is 'openalexR' [2]. The tool was developed by Aria (2022). This library provides a user interface to the OpenAlex APIs and supports the retrieval of bibliographic information about publications, authors, sources, institutions and concepts. The package can be installed either from CRAN or GitHub. The package mainly supports the following four functions for data retrieval:

1. `oa_query()`: A query is generated from the arguments provided by the user.
2. `oa_request()`: Once the appropriate query is generated through oa_query or manually input by the user, the set of entities matching the query can be downloaded with the help of oa_request. A list of JSON objects is returned.
3. `oa2df()`: converts the JSON object to a data frame or tibble.
4. `oa_fetch()`: All three functions above can be executed in a single step using oa_fetch.

For example, if the OpenAlex ID for a particular author or work is available, the metadata related to it can easily be retrieved with the help of oa_fetch. For more information, one can refer to the documentation provided.

4.1.3. rcrossref

rcrossref [3] provides an R interface to retrieve data through the Crossref APIs. The package was developed by Chamberlain et al. (2021) as part of the rOpenSci set of packages. It can be installed via CRAN or Github. Once the installation is done, it can be imported and used accordingly in R. It ensures that

[2] https://github.com/massimoaria/openalexR

[3] https://github.com/ropensci/rcrossref

all the etiquette needed to access data from the Crossref database is followed. The polite API pool provides improved performance and consistent response times. To access the polite pool, one must provide their email ID along with the request sent. Using the package, one can store their email ID as an environment variable in the '.Renviron' file and then load the package and request the required metadata.

Metadata related to works in a journal can be retrieved by using the function `cr_journals()`. It takes either the ISSN of the journal or keywords and fetches information about works published in the journal. Additionally, one can use the function to retrieve metadata related to multiple DOIs. Similarly, the function `cr_works()` helps us to fetch metadata regarding a single work. It allows us to search by a single or a set of DOIs. Multiple filters like location, funder, ORCID etc., can also be added to the query. The output is returned in JSON format, and they recommend the use of a JSON viewer to view the output.

4.1.4. opencitingpy

opencitingpy [4] is a Python library that can be utilised to obtain data from the OCC. Developed by Saralegui (2021), this package allows easy data extraction via the OpenCitations API. The package can be downloaded via PyPi. It's easy to use and does not require memorising URLs or formatting the responses received from the APIs. The Python package can be easily installed via pip. The metadata related to a list of scientific articles can be fetched by providing their corresponding DOIs. One can access all the endpoints available in the OCC currently.

4.1.5. semanticscholar

semanticscholar [5] is a Semantic Scholar Academic Graph API client library written in Python which was developed by D. Silva (2022). It can be downloaded from Pypi and imported & used as needed. Metadata related to scientific papers or authors can be fetched with the help of the library. Access to the public API and S2 Data Partner's API (using a private key) is enabled. It mainly supports two functionalities

1. Author Lookup - an author can be looked up by providing the corresponding Semantic Scholar ID or using a keyword search.

2. Paper Lookup - A paper can be looked up by providing the corresponding ID, like DOI or using a keyword search.

Similarly, one can easily navigate through the results regardless of the number of pages. All fields are included in the response by default. By using query

[4] https://github.com/unaisaralegui/opencitingpy
[5] https://github.com/danielnsilva/semanticscholar

parameters, one can easily include only the required fields, such as the title and the publication year. The results can also be filtered according to query parameters like the publication year.

4.1.6. USGS BiblioSearch

The tools discussed previously are all designed to retrieve metadata from a single database, whereas the USGS BiblioSearch [6] tool can be used to retrieve metadata from multiple databases. It is a Python tool developed by Kleist and Enns (2022), which can be used to search, clean, and compile literature citations from different citation databases. The tool can retrieve citation information from the WoS, Scopus, USGS Pubs Warehouse, and ScienceBase databases. It makes use of APIs and other web services to perform a systematic search of the different citation databases. It then cleans and compiles the result obtained.

Different levels of search results can be retrieved after a search operation, i.e., filtered results from across databases, raw results from each database, detailed record of a search query and summary of search results. Since the WoS and Scopus databases are proprietary databases, the users must have access to these databases to retrieve the metadata. On the other hand, both USGS Pubs Warehouse and ScienceBase provide open access.

4.2. Drawbacks of the existing Packages or Libraries

Even though many libraries are available for retrieving citation data across different citation databases, only a few packages or libraries enable the retrieval of data from multiple citation databases. For example, the Pybliometrics package can only retrieve the citation data from the Scopus database. Similarly, rcrossref can only retrieve citation data from the Crossref database. One major challenge with extracting citation data from different databases is the absence of a uniform interface. This can make it difficult to extract data consistently and accurately, as each database may have its own protocol for the extraction of citation data. As a result, researchers may need to spend more time and effort to extract data from each individual database.

Another significant challenge is that the output or citation data retrieved may be in different formats. This can make it difficult to analyse the data consistently, as different formats may require different approaches to processing and analysis. Additionally, researchers may need to spend more time and effort standardizing the data formats across different databases, which can be time-consuming and may require technical expertise. Without a consistent format for the data, it can also be difficult to compare and contrast citation data from

[6] https://code.usgs.gov/fort/bibliosearch/-/tree/v1.0.0

different data sources. Moreover, there is a possibility that these tools may be written in different programming languages. Researchers may need to spend additional time and effort learning multiple programming languages. This can be a particular challenge for researchers who do not have a strong background in programming, as they may find it difficult to navigate the technical aspects of using these tools.

Usually, in bibliometric studies, there is a heavy focus on the citation counts between entities like journals or authors. In most cases, the existing tools available simply retrieve data related to sets of journals or authors. These tools rarely offer an option for aggregating citation counts between a set of entities like journals or authors for a specific time frame.

4.3. Advantages of CitaTrack

Most of the drawbacks of the packages mentioned above are addressed satisfactorily with the implementation of the new package 'CitaTrack'. The package is designed to be user-friendly and provides a uniform interface to access the citation data from five different citation databases, Scopus, OpenAlex, OCC, Semantic Scholar and Crossref. An example of a citation analysis that was performed is described in detail in Chapter 6.

Currently, only five data sources are supported, but the package is scalable and can accommodate more databases if necessary. It is designed to improve the flexibility of the data retrieval process, that is, it offers the user the possibility to choose only the required databases.

Also, the package eliminates the need to go through the documentation of each and every citation database separately. In addition, the tool can be used by people with little technical knowledge of HTTP requests or APIs. As a result, users may save a lot of time and resources.

The citation counts between the works published in a set of journals or by a set of authors in a particular timeframe are aggregated and provided as output to the user. The output obtained is in a citation matrix format (a table of cross-citations), making it easier to analyze, draw insights from the data and can be used for further post-processing. For example, the ranking of journals can be performed with the help of the cross-citation table obtained using the package.

4.4. Comparison between the tools

Table 4.1 illustrates the comparison between the tools and highlights their key features. The tools are compared based on several factors. These factors include the programming language used to develop the tool, the repository from which the tool can be downloaded, whether the tool can extract data from

Table 4.1.: *Comparison between the tools used for data extraction*

Tools	Language	Repository	Single/ Multi Source	Data source	Entity	Authenti cation	Output Format	Cache
Pybliometrics	Python	PyPI/ GitHub	Single	Scopus	works, authors, affiliations, serial titles	yes	xml/ JSON	Yes
openAlexR	R	CRAN/ GitHub	Single	OpenAlex	works, authors, sources, institution, concepts	no	tibble/data frame/ JSON	No
rcrossref	R	CRAN/ GitHub	Single	Crossref	works, funders, members, prefixes, types, journals	yes	bibtex/CSL-JSON /RDF-XML	No
opencitingpy	Python	PyPI/ GitHub	Single	OCC	works	no	text	No
semantic scholar	Python	PyPI/ GitHub	Single	Semantic Scholar	works, authors	yes	text/ JSON	No
USGS BiblioSearch	Python	GitHub	Muti	WoS, Scopus, USGS publications warehouse, USGS sciencebase	works	yes	spreadsheet/ raw compiled results	No
CitaTrack	Python	GitLab	Multi	Crossref, OpenAlex, OCC, Scopus and Semantic Scholar	works, authors	yes	text/ CSV	Yes

single or multiple sources, the data sources supported by the tool, the entities that can be extracted using the tool, whether the tool provides authentication mechanisms, the output format of the retrieved data, and whether the tool caches results. By considering these factors, we can gain a better understanding of the capabilities of the tools that can be used for citation data extraction.

5. Journal Ranking

Journal ranking is performed in academia by various organisations, institutions, and individuals to assess a journal's relevance, impact, and quality. In this chapter, we will examine some of the widely used citation-based metrics for ranking journals. However, Bradley-Terry, one of the statistical models, is different from the other citation-based metrics. It is a powerful statistical model that can be used to rank journals based on paired comparisons. Additionally, we will talk about the potential advantages of the Bradley-Terry model compared to other ranking approaches.

5.1. Background

Journal ranking plays a crucial role in academic research as it helps researchers and institutions to make informed decisions about where to publish and which journals to follow. Top-ranking journals often exercise their discretion in selecting academic papers and have set standards for choosing an article for publication. To maintain their position among other journals, they are highly selective and ensure that the contents of the academic papers published are of top quality. So, it is widely accepted that any work published in a top-ranking journal has a certain level of prestige and value associated with it. Moreover, the distribution of scientific works within the research community is assured when academic works are published in top-ranking journals. According to Chang et al. (2010), when suitable data is not available for the evaluation of an academic paper, its quality is often inferred from the reputation of the journal in which it is published.

As observed by Hudson (2013), the approaches commonly used to evaluate a journal's quality and impact are mainly grouped into two categories, namely, the subjective peer review approach and the objective approach based on citation metrics.

Subjective peer review approaches usually employ a panel of experts to evaluate the journals. The final ranking of journals is arrived at by aggregating the expert's opinions. The main merit of the method is that the journals are evaluated by a panel comprising expert scholars in the field (Bontis & Serenko, 2009). However, this approach can be expensive and time-consuming if the number of evaluators is high. Moreover, their opinion is subject to personal biases and may not reflect the true picture (Belter, 2015). The process is also manual and subject to limitations thereof. The ranking is subject to the knowledge of the experts in the panel and can also be prone to manipulation and external influences.

The objective approach is mainly based on citation metrics which in turn may be based on a formula (Hudson, 2013). The main strength of this method is that the disadvantages of the subjective peer review approach are eliminated. Most citation metrics are based on journal-level pairwise citation counts. As the calculation of citation counts is automated, the disadvantages of manual methods are removed. As the human intervention is minimal, the process is less error-prone. In objective approaches, a large number of journals can be included in the pool for processing and ranking and the processing time is much faster compared to the subjective peer review approaches. However, citation-based metrics have been widely criticised because they can fail to signify the impact or importance of academic work in a domain. The counted citations may include several works having a critical opinion about the cited work. The limitations of the objective approaches will be discussed in the upcoming section.

5.1.1. Citation-based Metrics

Citation-based metrics are often used to evaluate the scientific impact of research publications. These metrics assess the impact of a journal or a scholarly work by analyzing the number of times it has been cited by other researchers. Waltman (2016) remarks that these metrics provide information on the impact of scientific units such as individual or groups of researchers, research institutions or scientific journals, and it plays a prominent role in the evaluation of scientific research. Moreover, he also points out that the significance of citation impact indicators in research evaluation has increased a lot during the past few decades, as evidenced by the rapid growth of scientific literature on citation impact indicators. The simplest citation-based metric is based on citation is the total number of times a journal's articles have been cited, i.e., the citation count (Walters, 2017). However, a major limitation of this metric is that it is influenced by the size of the journal, as larger journals are likely to have more articles and, thus, more opportunities for citations.

Roldan-Valadez et al. (2019) highlight the classification of bibliometrics based on different contexts or perspectives, in particular, metrics based on the prestige of the citing journal and metrics which are based on normalised citation scores. The former is based on a methodology that assigns more weight to citations from esteemed journals. The first category includes metrics such as SCImago Journal Rank (SJR) by Scopus, as well as the Eigenfactor by WoS. On the other hand, according to the authors, the second group of metrics attempts to standardise the citation rate within a particular subject group and includes metrics like CiteScore, IF and SJR. A brief overview of the citation-based metrics is provided below:

Journal Impact Factor

Among the oldest and most widely used citation metrics is the Journal Impact Factor (JIF). It is the ratio of the total number of citations received by a particular journal and the total number of papers published in the journal in the previous two years, which is equivalent to the average citation rate per published item in a two-year time period (Garfield, 1972). It was developed in the 1950s and is based on the data retrieved from the WoS database. It is calculated on a yearly basis and is published in the Journal Citations Report[1] by Clarivate Analytics (Roldan-Valadez et al., 2019)

According to Garfield (1994), the JIF is calculated as follows:

$$T = \text{Total cites in a year } Y$$
$$C = \text{Cites in year } Y \text{ to articles published in } Y-1 \text{ and } Y-2 \text{ (} C \text{ is a subset of } T\text{)}$$
$$N = \text{Number of articles published in } Y-1 \text{ and } Y-2$$
$$I = C/N = \text{Journal Impact factor for the year } Y$$

Here, Y represents the current year for which the JIF is being calculated, $Y-1$ represents the year immediately prior to the current year Y and $Y-2$ represents the year two years prior to the current year Y.

Even though the JIF is widely used currently for ranking journals, originally the idea was conceived for a different purpose, specifically to aid libraries in determining which journals to index and acquire for their collections (Haustein & Larivière, 2014). The metric's designer was of the opinion that it was not appropriate to use the metric for the evaluation of works or individuals (Garfield, 1963). McKiernan et al. (2019) remark that the correlation between the JIF and the prestige of a journal has caused academics to aspire to publish their research in journals with high JIF scores.

The JIF can fluctuate depending on several factors, such as the number of authors involved in writing the paper, the language in which the paper was written and the length of the scientific publication (Selby, 2020). Few works receiving a higher number of citations can boost the JIF considerably. Further, the citations received by works in a particular journal are not uniformly distributed. Therefore, it cannot be considered an appropriate metric for measuring the impact of all the works in a particular journal. Also, it takes into consideration only the works published within a two year window (Bornmann & Williams, 2017). However, some works get noticed several years after their publication. A significant limitation of the IF is its reliance on self-citations.

[1] https://jcr.clarivate.com/jcr/home

The use of self-citation in the IF has been subject to criticism due to the potential bias it introduces by inflating the citation count (Yuen, 2018). Another criticism regarding the metric is that the numerator includes all citations to a journal, including those to non-research articles, which can inflate the citation count. Meanwhile, the denominator only includes 'citable' items (e.g., journal articles) (Walters, 2017).

Although there are various limitations associated with it, the IF continues to be widely used and accepted. The academic community and the publishers of the journals continue to rely on this metric for the ranking and evaluation of journals, works or individuals.

h-index

The h-index, a metric to measure the impact of a researcher's scientific work, was developed by Jorge Hirsch in the year 2005. The metric is freely accessible through Scopus and Google Scholar (Grech & Rizk, 2018).

According to Hirsch (2005), "*a scientist has index h if h of his or her Np papers have at least h citations each and the other (Np-h) papers have less than or equal to h citations each*".

The metric was designed to measure "the broad impact of an individual's work or 'overall scientific impact' "(Barnes, 2017) and has become one of the significant metrics used to evaluate the productivity and impact of researchers (J. A. T. D. Silva & Dobránszki, 2018).

Several variations of the metric have been developed; however, none of the variants exceeded the metric's ability to measure the impact of researchers. The main characteristics, as observed by Koltun and Hafner (2021) are:

1. The research output of a scientist or researcher is encapsulated into a numerical value which in turn can be used for comparison or ranking of researchers.

2. The metric has the ability to quantify impact irrespective of the number of publications of a researcher or their career stage.

3. The calibration of thresholds or parameters is not required.

4. Although there have been criticisms, it is regarded as a reliable measure of a scientist's impact.

Setti (2013) noted some positive aspects of the h-index when compared to the JIF. Firstly, it is insensitive to a sudden increase in the number of non-cited papers or a few highly cited contributions, which may occur accidentally. Additionally, it combines the number of papers published and the citation rate, which reduces the apparent overperformance of small journals as measured by the JIF.

The metric's usage for the evaluation of individuals has differing opinions among the bibliometricians due to some of its limitations. As noted by Kreiner (2016), even if a work receives several thousand citations, the h-index can never go beyond the number of works published by a researcher. They also point out that self-citations are not excluded and can result in the manipulation of the metric. Further, the metric cannot be used for comparing researchers at different stages of their careers or individuals from different scientific domains. It does not differentiate between the authorship positions as well, so a noteworthy amount of work can be done by the first author and comparatively lesser work by the middle or last author (Grech & Rizk, 2018). However, there is no system in place to differentiate between the nature of work. The authors also note that the metric is dependent on the discipline of research since research that is extremely specialised may have a smaller audience, which results in fewer citations.

Eigenfactor

The Eigenfactor metric was developed by Bergstrom (2007) and the metric assesses the overall impact of a journal on the academic literature, that is, the collective worth contributed by all the articles published in the journal during a given year. The use of just the citation counts as a metric was heavily criticised for the reasons already discussed and the IF has been subject to many criticisms. Hence, Eigenfactor was developed as an alternative metric since IF is susceptible to manipulation and fails to account for the quality and prestige of the citing journals. The metadata used for ranking is fetched from the WoS database. The methodology used to rank journals is similar to Google's Page Rank Algorithm. Bergstrom (2007) describes the approach as follows: "we measure the importance of a citation by the influence of the citing journal divided by the total number of citations appearing in that journal".

Self-citations are not taken into account by Eigenfactor in its evaluation of a journal's impact, therefore, manipulating the metric is often difficult (Setti, 2013). Further, the metric is often positively correlated with journal size, meaning that larger journals tend to have higher scores (Pajić, 2015; Setti, 2013). It is calculated using citation data from a five-year period, and it can be calculated for a set of journals by adding together the individual scores of each journal in the group (Yuen, 2018).

Article Influence Score

The Article Influence Score (AIS) is independent of the size of the journal and is based on the Eigenfactor (Walters, 2017). Both the Eigenfactor score and the derived AIS were introduced to address the limitations of the IF (Varin et al., 2016). The metric determines the importance of a journal by assessing the influence of each article published in the journal. The following formula can be used to calculate the metric from the Eigenfactor Score(Clarivate, 2021):

$$\text{Article Influence Score} = \frac{0.01 \times \text{Eigenfactor Score}}{X}$$

Here, X is the ratio of the number of articles published by a journal over a five-year period to the total number of articles published by all journals during the same five-year period. The data from the WoS database is used for the calculation of this metric and is available freely at eigenfactor.org or in the Journal Citation Reports of Clarivate Analytics.

The Eigenfactor and the AIS have many common characteristics as the latter was derived from the former. They both use a five-year citation window to capture the impact of a journal (Pajić, 2015) and both exclude the self-citations during its calculation (Setti, 2013). Both take into account the quality and influence of the citations received by a journal, rather than just the raw number of citations.

SCImago Journal Rank (SJR)

SCImago Journal Rank [2] is a metric which is based on citation data extracted from the Scopus database. SJR employs a similar algorithm to the one used by Google Page Rank, whereby citations are weighted according to the perceived 'prestige' of the journal that made the citation (Yuen, 2018). SJR takes into account not only the number of citations received by a journal, but also the importance of the journals that cite it (Colledge et al., 2017). For instance, a citation from a highly ranked journal is given more weight than a citation from a lower-ranked journal.

Guerrero-Botea and Moya-Anegón (2012), who have contributed to the development of the SCImago Journal Rank, mention that the calculation of SJR is done in two phases, in the first phase, the prestige of each journal is calculated and this measure is size dependent. For more information on the calculation of the prestige, please refer to the paper. In the second phase, the normalisation of the measure obtained (in phase 1) is performed. The SJR uses a normalization method where the prestige gained by each journal (measured as PSJR2) is divided by the ratio of its citable documents relative to the total number of citable documents in the field to account for differences in journal size. According to the author, it can be expressed as:

$$SJR_i = \frac{PSJR2_i}{\left(Art_i / \sum_{j=1}^{N} Art_j\right)} = \frac{PSJR2_i}{Art_i} \cdot \sum_{j=1}^{N} Art_j$$

[2] https://www.scimagojr.com/journalrank.php

One of the differences between the JIF and the SCImago Journal Rank is that the publication window considered for the latter is three years, whereas, for the former, it is two (Roldan-Valadez et al., 2019). Further, citations are not equally weighted in the network, unlike the JIF, citations from the most prestigious or influential journals are weighted more heavily. To mitigate the potential impact of excessive self-citations on the evaluation of a journal's influence, the metric sets a maximum limit of 33% of self-citations that each journal can receive (Setti, 2013). This limit helps to prevent manipulation of the metric to a certain extent.

CiteScore

CiteScore [3], a relatively new citation metric by Elsevier B.V., was introduced in 2016 and is a competitor of the JIF. According to Zijlstra and McCullough (2016), the metric provides a transparent and up-to-date evaluation of a journal's influence and can aid one in determining where to submit their next publication. This metric is determined by dividing the total number of citations received by articles published in the current year by the total number of articles that the journal has published over the previous three years (Colledge et al., 2017). It can be written as (Roldan-Valadez et al., 2019):

$$\text{CiteScore} = \frac{C}{N}$$

where, C is the Citations received in the year Y to articles published in the years $Y-1$, $Y-2$ and $Y-3$ and N is the total number of articles published in the years $Y-1$, $Y-2$, and $Y-3$. The metric was developed to address the lack of simple citation-based journal metrics in Scopus. Only peer-reviewed publications are used for the CiteScore calculation, and all serial titles indexed in Scopus are included. It is free and is available to both subscribers and non-subscribers. Unlike the JIF, the metric offers transparency of underlying data and hence can be validated easily (McCullough, 2022). Furthermore, the Scopus metric for newly published serial titles is only available a year after they were indexed in the database.

Colledge et al. (2017) mention that the metric is a useful tool for assessing the citation impact of serial titles that belong to a common subject area. The utilisation of the metric in the assessment of journals pertaining to various fields such as molecular medicine, nursing, behavioural neuroscience, trade, and dermatology has been observed (Roldan-Valadez et al., 2019).

The main similarities and differences between the JIF and CiteScore are listed in Table 5.1. It is the publication window, transparency and coverage that mainly differentiate the two metrics, both of which are very similar. The

[3] https://www.scopus.com/sources

Table 5.1.: *Comparison between the JIF and CiteScore (J. A. T. D. Silva & Memon, 2017)*

Characteristics	JIF	CiteScore
	Similarities	
Calculation	Citation per document	Citation per document
Simplicity	Yes (calculates mean)	Yes (calculates mean)
Annual snapshot	Available for reporting purposes	Available for reporting purposes
Inclusion for error correction	No	No
Field-specific differences	Addressed	Addressed
Self-citation	Included	Included
	Differences	
Proprietary	Non-publisher	Publisher
Database used	WoS	Scopus
Publication window (years)	2	3
Numerator versus denominator	Inconsistent	Consistent
Document types included	Articles and reviews	All types, including articles, reviews, letters and editorials
Sources	Journals only	All (journal, conference proceedings)
Coverage	11,000 titles or 32,925 journals	22,000 titles or 22,256 journals
Transparency	Subscription-based	Freely available (except for in-depth analysis)
Availability of updates	Yearly	Monthly
Effect of editorial policy	Influenced	Not influenced

journals having a high JIF can have a lower CiteScore owing to the difference in coverage of the databases. Noorden (2016) observes that documents like editorials, letters, news items etc., are counted by the new metric, and these documents are less cited by researchers or scholars; therefore, the average is often dragged down.

5.1.2. Altmetrics

Apart from the commonly used subjective peer review and the objective citation metric-based approaches for measuring the impact of the journals, alternative metrics/altmetrics are emerging as a novel option to measure the societal impact of research output. Altmetrics take into consideration social media activity and also other forms of significant research output that is not typically included in traditional peer-reviewed publications (Williams, 2017). For instance, in 'article-level metrics', the data from social media such as downloads, recommends, clicks, notes, tweets, shares, views, likes, saves, posts, tags, bookmarks, discussions, trackbacks, and comments are counted to assess the impact, instead of the citation counts between papers (Bornmann, 2014). In comparison to traditional metrics, four main advantages were identified by the author:

1. Broadness: altmetrics measures societal impact, in other words, it can measure impact beyond science.

2. Diversity: The impact of all scholarly products (such as data sets, software and patents) can be measured, not only academic papers.

3. Speed: The other metrics, like the IF, take longer to be published, while the altmetrics enable the measurement of the impact of a scholarly product within a short span of time after its release.

4. Openness: In altmetrics, data collection is not an issue as it can be collected without any difficulty.

Despite having many advantages, altmetrics have several disadvantages, as observed by Thelwall (2020). First, there is no direct and easy method for data collection, and it is also time-consuming. The other disadvantages mentioned by the author are low coverage, difficulty with field normalisation, incomplete and biased coverage of impact areas, incomplete coverage of impact types and lack of quality control. Also, compared to traditional bibliometrics, altmetrics are more susceptible to manipulation (Bajwa & Mehdiratta, 2021). For instance, Twitter mentions can be created by bots or fake accounts. Moreover, there is no formal process for linking user's online profiles to their offline identities on social media websites. In cases where citations alone are insufficient for evaluation, such as when evaluating non-academic impacts or nonstandard outputs, the altmetrics can prove to be very useful (Thelwall, 2020).

Walters (2017) mentions that altmetrics have not gained widespread acceptance as replacements for established measures of a journal's impact and reputation despite their potential. Williams (2017), on the other hand, highlights that, altmetrics are not meant to replace traditional citation-based measurements but rather to provide additional and complementary information.

5.1.3. The Stigler Model

The Bradley-Terry model (Bradley & Terry, 1952) is a quantitative approach based on pairwise comparisons that can be used to rank journals. A model based on paired comparisons compares each item in a set against every other item, and the final ranking is determined by the relative preferences or rankings of each item within the set. Bradley-Terry models are commonly used in fields such as sports and statistics (Liner & Amin, 2004). The model assumes that the relative strength of preference between two items can be represented as the odds ratio of one item being preferred over the other. To quote Turner and Firth (2012), the basic assumption of the model is that, " in a 'contest' between any two 'players', say player i and player j, $(i, j \in \{1, ..., K\})$, the odds that i beats j are α_i/α_j, where α_i and α_j are positive-valued parameters which might be thought of as representing 'ability'."

The Stigler model, which utilises a stochastic approach to model a matrix of cross-citation counts, is an example of the Bradley-Terry model (Varin et al., 2016). The model can typically be fitted using maximum likelihood estimation. The basic idea is to find the set of model parameters that maximise the likelihood of observing the data given the model. Stigler (1994) looked at citations as import-export statistics reflecting intellectual influence. In his words, "when one journal, say Journal A, prints a paper containing a citation to work previously published in another, Journal B, we may consider this as indicating an instance of the export of intellectual influence from Journal B and the import of intellectual influence by Journal A" subject to certain limitations.

The journal that publishes the original article is considered the source journal that exports the bibliographic reference, while the journal that references the original work is regarded as the recipient that imports it. The model was designed to measure the 'export scores' for academic journals (Selby, 2020). The export score could be considered as a measure of one journal's influence on other journals. This score is calculated relative to a baseline journal that is chosen arbitrarily.

According to Varin et al. (2016), "the log-odds that journal i exports to journal j rather than vice versa are equal to the difference in the journal's export scores".

$$\text{log-odds(journal } i \text{ is cited by journal } j) = \mu_i - \mu_j$$

In this equation, μ_i is journal i's export score. The author notes that, greater the export score, the greater the likelihood of exporting intellectual influence. A journal with a high export score is considered more influential than one with a low score; therefore, a ranking can be derived based on the export scores of a set of journals. An example of ranking journals using the Stigler model is given in Chapter 7.

5.1.4. Advantages of Stigler's model over other metrics

The model exhibits several advantages according to Varin et al. (2016):

1. A journal's size is insignificant, and there's no need to perform normalisation based on journal size, unlike the other metrics.
2. Journal self-citations are ignored, which is especially beneficial where the number of citations is manipulated through self-citations.
3. Only the citations among the journals included in the comparison set are considered.
4. The prestige of the citing journal is taken into consideration.

There are some disadvantages to the model as well. For instance, the model fails when the citation network is disconnected. Liner and Amin (2004) note that it may not perform well for a group of journals that do not exhibit a strong bilateral trade in citations. In such scenarios, there may be a lack-of-fit problem, which could compromise the reliability of the model in terms of predicting the propensity to export information between journals. Additionally, the traditional model Bradley-Terry model cannot process fractional citations. However, the advantages of the model outweigh the disadvantages.

6. A Motivating Example: Part A - Citation Analysis

Research in economics frequently involves the use of bibliometric analysis methods (e.g. Ketzler & Zimmermann, 2012; Wei, 2018). The analysis of citations can help identify influential research, track trends over time, evaluate the impact of research, and identify potential collaborators.

In Chapter 4, we showed examples of tools available for extracting citation data from different citation databases. As we saw, there is a need for more tools that can extract citation information from multiple databases. We developed CitaTrack, a Python package for extracting citation data from 5 different citation databases to address this issue. It is written using Python 3, and API wrappers were developed for the respective APIs of the data sources as part of the package. The citation databases are accessed by the tool to extract information about citations between authors or journals. The package is designed to be simple and easy to use. It provides a uniform interface to extract citation data related to journals and authors from five citation data sources: Scopus, OCC in conjunction with Crossref, OpenAlex, Crossref and Semantic Scholar in conjunction with Crossref. For better performance, the metadata is cached after it is extracted. More information about the tool and its implementation is discussed in Chapter 8.

Even though several studies have been conducted in the field of citation analysis over the years, studies covering newer databases like Semantic Scholar and OpenAlex are few in number. We will now look into an example of an analysis and journal ranking that was conducted using the tool. This example is divided into two parts: citation analysis and modelling. CitaTrack was used to extract data from the five data sources for the articles published in the top twenty economic journals in the time period 2016-2020 for both citation analysis and modelling (ranking of journals). The journals were selected from Google Scholar's list [1] of the top 20 journals in the field of Economics. Google Scholar ranks the journals based on the h-index. The titles of the journals chosen and their abbreviations are given in Table 6.1. The five-year citation window was chosen because it results in consistent rankings of journals over an extended period of time (Pajić, 2015). According to Setti (2013), "a five-year citation window reduces fluctuations between years and better reflects the impact of papers in most disciplines". The focus of this chapter will be

[1] https://scholar.google.de/citations?view_op=top_venues&hl=en&vq=bus_economics

on citation analysis. The results of this analysis illustrate the variability in data from different sources and skewed distributions of citations across the top 5 journals in economics. With this, we will show why using multiple data sources to perform journal ranking is beneficial to gain a more comprehensive understanding of the ranking of journals. We also demonstrate why indicators using the weighted average of citations, such as IF, may not be a reliable measure for ranking journals.

Economic journals have become more significant in recent years as the world has become increasingly interconnected. In the wake of globalisation and increased levels of international trade caused by the liberalisation of policies worldwide, these journals have become a valuable source of information for economists, business houses and academics alike. They offer a deep insight into the latest economic trends, developments and issues. They also shed light on the data and analysis that influences economic decision-making and the framing of policies. Additionally, these journals provide a platform for economists to share their research with the public.

Table 6.1.: *Selected economic journals along with their abbreviations*

SNo.	Journal Name	Abbreviation
1	American Economic Review	AER
2	Review of Financial Studies	RFS
3	Quarterly Journal of Economics	QJE
4	Journal of Political Economy	JPE
5	Journal of Finance	JF
6	Review of Economic Studies	RES
7	Econometrica	Eco
8	Journal of Economic Perspectives	JEP
9	Journal of Public Economics	JPEco
10	Review of Economics and Statistics	REStat
11	Journal of Development Economics	JDE
12	Economic Journal	EJ
13	Journal of Monetary Economics	JME
14	Journal of International Economics	JIE
15	Economic Modelling	EM
16	Economics Letters	EL
17	Journal of the European Economic Association	JEEA
18	Journal of Economic Literature	JEL
19	American Economic Journal: Macroeconomics	AEJM
20	European Economic Review	EER

There is a broad consensus that the top 5 economic journals are Econometrica, the Journal of Political Economy, the Quarterly Journal of Economics, the American Economic Review and the Review of Economic Studies (Heckman &

Moktan, 2020; Kalaitzidakis et al., 2011). These journals are highly esteemed in the academic world and hold significant influence. Heckman and Moktan (2020) observe that these journals are listed as the top five journals based on 'aggregate proxies of journal influence'.

6.1. Analyzing citations with the tool

A wide range of citation analysis use cases are addressed by the tool. Some use cases are citation count analysis, citation network analysis, co-citation analysis and trend analysis of citation counts. In citation count analysis, citation counts are used for assessing a publication's impact and significance. It is primarily used to evaluate journals and papers (Qiu et al., 2017). Citation network analysis allows one to identify and interpret the relationships between entities in a discipline based on their citations. McLaren and Bruner (2022) define it as a "review method that seeks to map the scientific structure of a field of research as a function of citation practices". Co-citation analysis is another practical application that helps identify the most frequently co-cited papers and authors in a field and helps understand key concepts and themes. As observed by Surwase et al. (2011), co-citation analysis "involves tracking pairs of papers that are cited together in the source articles". A similar approach can be applied to authors as well. One can easily identify the relationships between authors and their research areas (Kim et al., 2016). Moreover, the tool can be used to track citation trends over a period of time. CSV files generated as the output after the data extraction and processing via the tool can be used for this purpose. The two CSV files obtained as the tool's output are the file with cross-citation information in a tabular format and the one with detailed information about 'citing' and 'cited' works/authors. These two files can be used to perform all the analyses mentioned above.

The tool also supports analysis of the distribution of works across different data sources and comparative study among the data sources. In the upcoming sections, the distribution of works in the top 20 economic journals, along with the distribution of citations in the top 5 economic journals, are presented. As Selby (2020) points out, "many reviews in bibliometric literature of citation databases are concerned with paper- or author-level citation counts, or the number of documents indexed based on a subject, as a measure of 'coverage'". Further, the analysis of data from multiple sources provides us with valuable insights into the field of economics.

6.2. Distribution of works in economic journals

Figure 6.1 is a graphical representation of the number of works indexed in the 5 data sources of CitaTrack published between 2016-2020 for each of the top 20 economic journals. Data regarding the number of works for each journal was retrieved as part of the journal ranking discussed in Chapter 7. The

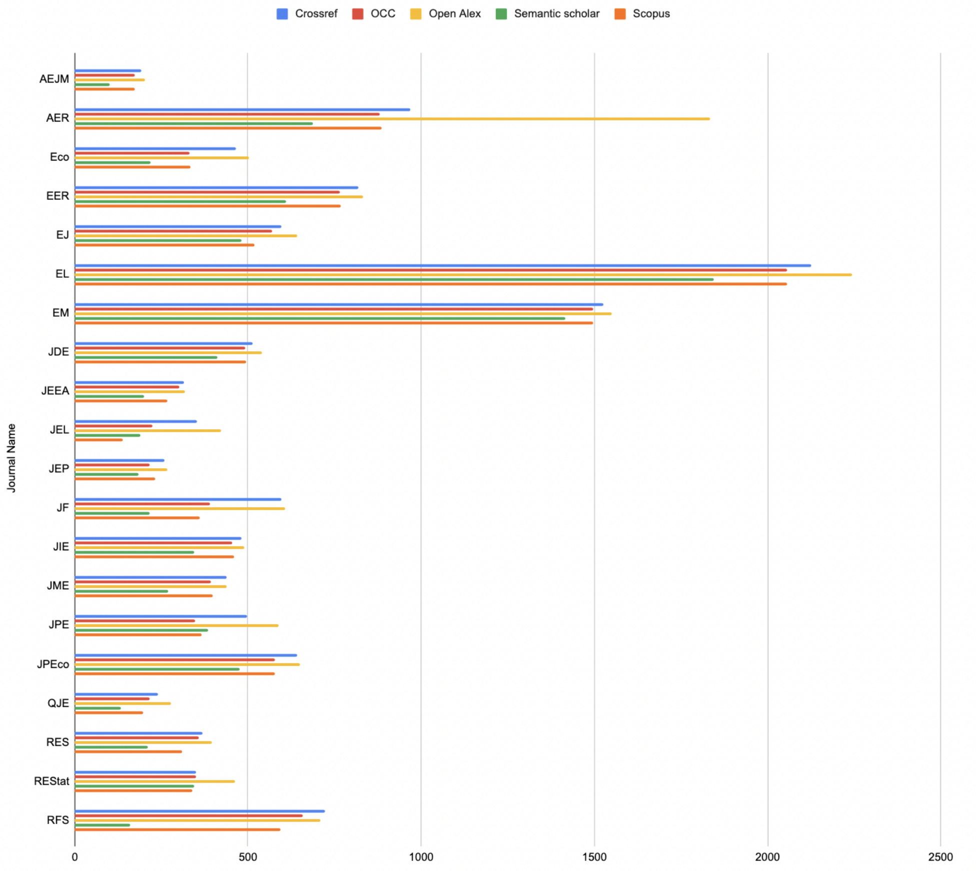
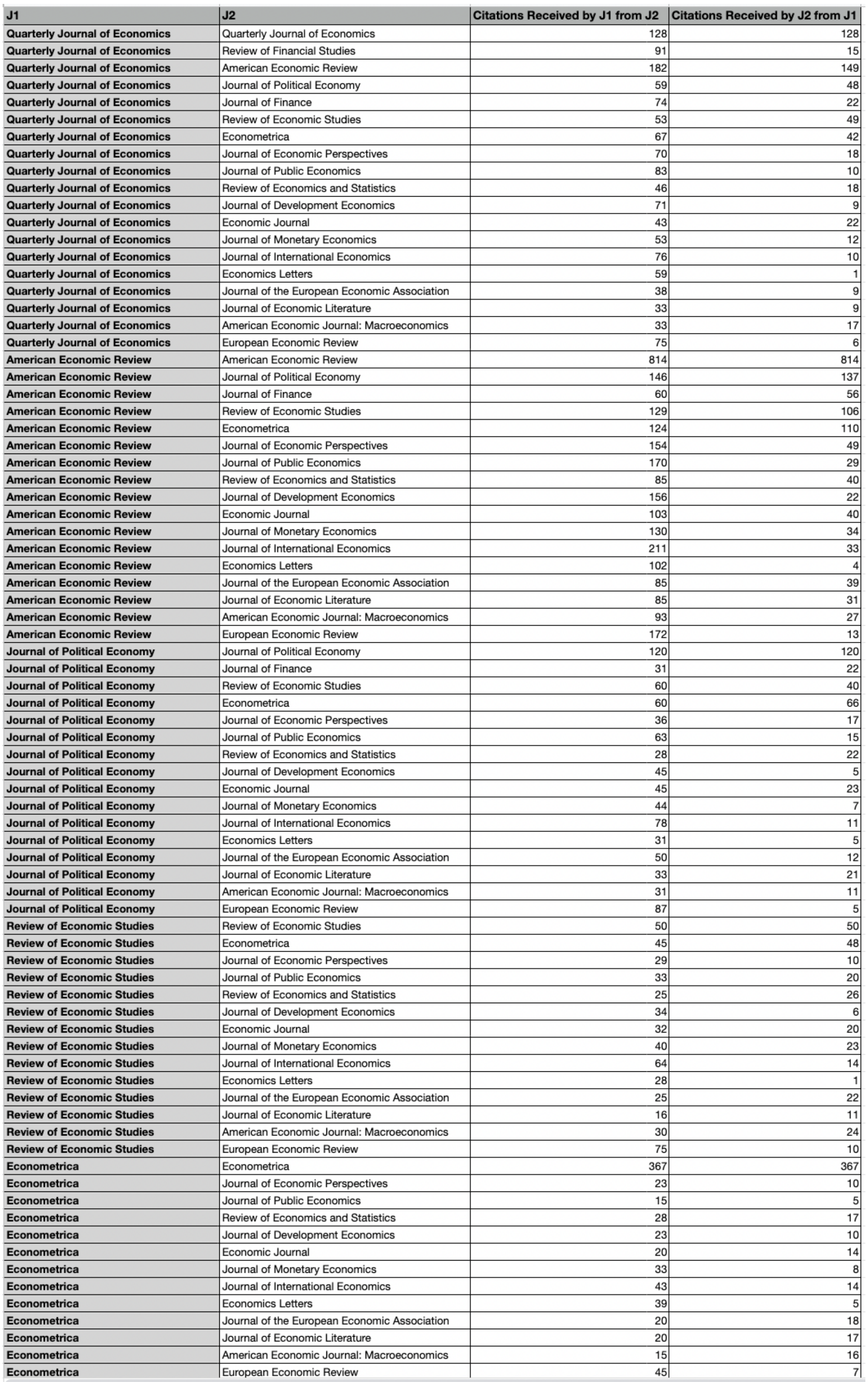


Figure 6.1.: *Distribution of works in 20 economic journals based on the data obtained from OpenAlex, Crossref, Scopus, OCC and Semantic Scholar*

retrieval of cross-citation data using CitaTrack is discussed in detail in Appendix A.1. In the graph, each work is grouped by its journal of publication. The x-axis of the graph depicts the number of works and the y-axis of the graph depicts the journal names. As can be seen from the figure, the data sources are distinguished by colour coding. The highest number of works is indexed by OpenAlex(13971), followed by Crossref (12468), Scopus (10967), OCC (10902), and the lowest by Semantic Scholar(8891). The figure clearly shows that the number of works indexed by Semantic Scholar is considerably lower than those indexed by other data sources. The number of works indexed by OpenAlex and Crossref is largely comparable, and that of Scopus and OCC is almost similar, with the latter having slightly fewer works than the former. OpenAlex is majorly based on Microsoft Academic data, and in the study conducted by Martín-Martín et al. (2021), it was found that Microsoft Academic has higher coverage than Scopus. Therefore, the current results align with the findings of Martín-Martín et al. (2021).

Of all the sources used, Semantic Scholar covers the least amount of publi-



J1	J2	Citations Received by J1 from J2	Citations Received by J2 from J1
Quarterly Journal of Economics	Quarterly Journal of Economics	128	128
Quarterly Journal of Economics	Review of Financial Studies	91	15
Quarterly Journal of Economics	American Economic Review	182	149
Quarterly Journal of Economics	Journal of Political Economy	59	48
Quarterly Journal of Economics	Journal of Finance	74	22
Quarterly Journal of Economics	Review of Economic Studies	53	49
Quarterly Journal of Economics	Econometrica	67	42
Quarterly Journal of Economics	Journal of Economic Perspectives	70	18
Quarterly Journal of Economics	Journal of Public Economics	83	10
Quarterly Journal of Economics	Review of Economics and Statistics	46	18
Quarterly Journal of Economics	Journal of Development Economics	71	9
Quarterly Journal of Economics	Economic Journal	43	22
Quarterly Journal of Economics	Journal of Monetary Economics	53	12
Quarterly Journal of Economics	Journal of International Economics	76	10
Quarterly Journal of Economics	Economics Letters	59	1
Quarterly Journal of Economics	Journal of the European Economic Association	38	9
Quarterly Journal of Economics	Journal of Economic Literature	33	9
Quarterly Journal of Economics	American Economic Journal: Macroeconomics	33	17
Quarterly Journal of Economics	European Economic Review	75	6
American Economic Review	American Economic Review	814	814
American Economic Review	Journal of Political Economy	146	137
American Economic Review	Journal of Finance	60	56
American Economic Review	Review of Economic Studies	129	106
American Economic Review	Econometrica	124	110
American Economic Review	Journal of Economic Perspectives	154	49
American Economic Review	Journal of Public Economics	170	29
American Economic Review	Review of Economics and Statistics	85	40
American Economic Review	Journal of Development Economics	156	22
American Economic Review	Economic Journal	103	40
American Economic Review	Journal of Monetary Economics	130	34
American Economic Review	Journal of International Economics	211	33
American Economic Review	Economics Letters	102	4
American Economic Review	Journal of the European Economic Association	85	39
American Economic Review	Journal of Economic Literature	85	31
American Economic Review	American Economic Journal: Macroeconomics	93	27
American Economic Review	European Economic Review	172	13
Journal of Political Economy	Journal of Political Economy	120	120
Journal of Political Economy	Journal of Finance	31	22
Journal of Political Economy	Review of Economic Studies	60	40
Journal of Political Economy	Econometrica	60	66
Journal of Political Economy	Journal of Economic Perspectives	36	17
Journal of Political Economy	Journal of Public Economics	63	15
Journal of Political Economy	Review of Economics and Statistics	28	22
Journal of Political Economy	Journal of Development Economics	45	5
Journal of Political Economy	Economic Journal	45	23
Journal of Political Economy	Journal of Monetary Economics	44	7
Journal of Political Economy	Journal of International Economics	78	11
Journal of Political Economy	Economics Letters	31	5
Journal of Political Economy	Journal of the European Economic Association	50	12
Journal of Political Economy	Journal of Economic Literature	33	21
Journal of Political Economy	American Economic Journal: Macroeconomics	31	11
Journal of Political Economy	European Economic Review	87	5
Review of Economic Studies	Review of Economic Studies	50	50
Review of Economic Studies	Econometrica	45	48
Review of Economic Studies	Journal of Economic Perspectives	29	10
Review of Economic Studies	Journal of Public Economics	33	20
Review of Economic Studies	Review of Economics and Statistics	25	26
Review of Economic Studies	Journal of Development Economics	34	6
Review of Economic Studies	Economic Journal	32	20
Review of Economic Studies	Journal of Monetary Economics	40	23
Review of Economic Studies	Journal of International Economics	64	14
Review of Economic Studies	Economics Letters	28	1
Review of Economic Studies	Journal of the European Economic Association	25	22
Review of Economic Studies	Journal of Economic Literature	16	11
Review of Economic Studies	American Economic Journal: Macroeconomics	30	24
Review of Economic Studies	European Economic Review	75	10
Econometrica	Econometrica	367	367
Econometrica	Journal of Economic Perspectives	23	10
Econometrica	Journal of Public Economics	15	5
Econometrica	Review of Economics and Statistics	28	17
Econometrica	Journal of Development Economics	23	10
Econometrica	Economic Journal	20	14
Econometrica	Journal of Monetary Economics	33	8
Econometrica	Journal of International Economics	43	14
Econometrica	Economics Letters	39	5
Econometrica	Journal of the European Economic Association	20	18
Econometrica	Journal of Economic Literature	20	17
Econometrica	American Economic Journal: Macroeconomics	15	16
Econometrica	European Economic Review	45	7

Figure 6.2.: *The total number of citations received by the top five journals within the set of 20 economic journals using CitaTrack*

cations in Economics. The result is in line with that of Arum (2016). As mentioned in section 3.3, the author remarked that the number of works indexed by Semantic Scholar is comparatively lesser and is limited to computer science and neuroscience journals. However, with the latest results, it is encouraging to see that journals from the field of economics are also indexed, even though the coverage of publications indexed is low. Further, the number of works published in AER indexed by OpenAlex is almost double than the ones indexed by other sources. This could be due to a data issue or data duplication, as it is a recently developed data source. Alternatively, it could simply be due to a better publication coverage. A detailed analysis could not be done due to time constraints.

Table 6.2.: *The percentage of works indexed in each data source for the top 5 economic journals*

	Data Source				
Journal Name	Crossref	OCC	OpenAlex	Semantic Scholar	Scopus
AER	7.77	8.07	13.12	7.72	8.07
JPE	3.99	3.18	4.21	4.31	3.33
Eco	3.72	3.03	3.60	2.46	3.04
RES	2.95	3.29	2.82	2.38	2.83
QJE	1.91	1.98	1.98	1.50	1.79

For each of the top 5 journals, Table 6.2 shows the share (percentage) of works indexed in each data source. According to the table, the highest number of works were published by AER (7.72% to 13.12%), followed by JPE (3.18% to 4.31%), Eco(2.46% to 3.72%), RES(2.38% to 3.29%) and QJE (1.5% to 1.98%). The lowest number of works were published by QJE, making it the most selective journal, and AER is the least selective among the group. The total number of publications in the period 2016-2020 is significantly smaller (approx. 2%) for the top five journals compared to the other 15 journals. Similar results were found in a study by Card and DellaVigna (2013). They found that QJE was the most selective, followed by JPE and RES, while Eco and AER were the least selective. These journals tend to be more selective in their publication standards, and they tend to prioritise articles that are expected to have a significant impact on the field. Therefore, a publication in one of the top five journals is considered prestigious.

6.3. Analysis of works in the top 5 economic journals

The number of citations a journal article receives can be seen as an indicator of its relative importance in the field. Citation analysis was performed to identify the papers (from the chosen journals) receiving the highest number of citations. A QJE article titled ‘Measuring Economic Policy Uncertainty’ was

Table 6.3.: *Top 1% most cited works in each of the top 5 economic journals*

Paper	Journal	Scopus ID	Citations
Measuring Economic Policy Uncertainty	QJE	2-s2.0-84997832043	84
Learning from inflation experiences	QJE	2-s2.0-84960353870	48
The surprisingly swift decline of US manufacturing employment	AER	2-s2.0-84978832070	61
The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment	AER	2-s2.0-84962800101	53
Railroads of the Raj: Estimating the impact of transportation infrastructure	AER	2-s2.0-85045219480	43
The power of forward guidance revisited	AER	2-s2.0-84991818004	41
The determinants and welfare implications of US Worker's diverging location choices by skill: 1980-2000	AER	2-s2.0-84960914080	38
The margins of global sourcing: Theory and evidence from US firms	AER	2-s2.0-85026858937	33
Liquidity trap and excessive leverage	AER	2-s2.0-84960903805	32
CoVaR	AER	2-s2.0-84978880727	32
Consumption inequality and family labor supply	AER	2-s2.0-84960157895	31
Downward nominal wage rigidity, currency pegs, and involuntary unemployment	JPE	2-s2.0-84988720343	46
Government spending multipliers in good times and in bad: Evidence from US historical data	JPE	2-s2.0-85043534657	35
The macroeconomic effects of housing wealth, housing finance, and limited risk sharing in general equilibrium	JPE	2-s2.0-85009963966	31
The pass-through of sovereign risk	JPE	2-s2.0-84979520745	30
Trade induced technical change? The impact of chinese imports on innovation, IT and productivity	RES	2-s2.0-84959117320	42
Trade and inequality: From theory to estimation	RES	2-s2.0-85014087187	26
Voting to tell others	RES	2-s2.0-85014186629	24
Prices, Markups, and Trade Reform	ECO	2-s2.0-84961252962	72
Uncertainty Shocks in a Model of Effective Demand	ECO	2-s2.0-85018298203	29
Really Uncertain Business Cycles	ECO	2-s2.0-85048007602	27

found to be the most cited across all data sources in the years 2016-2020. It was followed by the article “Prices, Markups, and Trade Reform”, published by Econometrica.

In this study, we attempted to investigate the percentage of citations received by the top 1% of works in the top 5 journals using the data from Scopus. With this, we intend to show the skewness in the distribution of citations received by a few works, specifically, 1% of the total number of works published in each of the chosen journals. The journals were chosen because they received more citations than the other journals. Scopus was chosen as the data source because the ranking obtained in Chapter 7 aligns with the general consensus on the top five economic journals. The total number of works indexed by the Scopus database for the set of 20 journals in the specific time period is 10967. The total number of works retrieved via CitaTrack published in QJE, Eco, JPE, RES and AER is 197, 334, 366, 311 and 886, respectively. Table 6.3 illustrates the top 1% of works for each journal having the most citations. Moreover, the total number of citations received and given out by the top 5 journals are listed in Fig 6.2.

Table 6.4.: *Distribution of citations among the top 1% of works in the top five economic journals*

Journal Name	Number of top 1% works	Total number of works	Percentage of citations
QJE	2	197	10 %
JPE	4	366	13.8%
Eco	3	334	13.4%
RES	3	311	12.7%
AER	9	886	12.3%

Using the data listed in Table 6.3 and Fig 6.2, the percentage of citations received by each journal’s top 1% of works can be calculated. For instance, the top 1% of works published in QJE, specifically just 2 out of 197 papers, received almost 10% of total citations received by the journal. Similarly, 9 works in AER, 4 works in JPE, 3 works each in RES and ECO (the top 1% of works) received 12.3%, 13.6%, 12.7% and 13.4% of the total citations (including self-citations) received by the particular journal. This is illustrated in Table 6.4. Clearly, the citation distributions in top journals are highly skewed, as demonstrated by this example. Several similar studies have demonstrated the skewed nature of journal citations in the past (Stern, 2013; Wall, 2009).

7. A Motivating Example: Part B - Modelling

The field of economics is continually evolving due to advances in research and development. Therefore, the literature on ranking journals in economics has grown significantly in the last 20 years (e.g. Bornmann et al., 2017; Hudson, 2013; Ritzberger, 2008), and it has been used as a tool to judge the research performance of economics departments (Kalaitzidakis et al., 2011). It is also observed that these journals play a crucial role in retaining old faculty and attracting faculty new faculty as well as talented graduate students to highly-ranked institutions. In a market where so many academic publications are available, they provide 'objective' information about the quality of those publications (Ritzberger, 2008).

This chapter focuses on our attempt to rank 20 economic journals for a period of five years, from 2016 to 2020, based on cross-citation data retrieved from five citation databases - OpenAlex, Crossref, OCC, Semantic Scholar and Scopus using the newly developed tool CitaTrack. In addition, we aim to verify that the top 5 journals in the ranking are in alignment with the general consensus on the topic.

As a first step in ranking journals, citation data was extracted using CitaTrack for the top 20 journals [1] in Economics. The chosen time window for the citation analysis performed using CitaTrack is five years, from 2016 to 2020. 'Ability scores' of individual journals were obtained using the Stigler model based on the data from the cross-citation table generated via CitaTrack. The set of journals was then ranked based on their ability scores. Further, the process was repeated for each cross-citation table pertaining to each data source. Even though several metrics are available for ranking journals, very few quantify the uncertainty associated with the rankings. Contrary to this, the model we have used for ranking, namely, the Stigler model, quantifies the uncertainty. A maximum likelihood method is used for fitting the Stigler model, and the uncertainty related to the ranking is estimated using quasi-variances. We'll examine it in more detail in the next section.

[1] https://scholar.google.de/citations?view_op=top_venues&hl=en&vq=bus_economics

7.1. Modelling

The Stigler model was discussed in Chapter 5 in detail. The procedure for ranking journals using the model based on the cross-citation data obtained via CitaTrack will be discussed in this section. The R package BradleyTerry2 (Turner & Firth, 2012) was used to obtain the rankings for economic journals.

To rank the journals, the cross-citation table obtained via CitaTrack was modified slightly to the appropriate format, i.e., pairs of journals with their corresponding frequencies of 'wins'. The self-citations counts are not taken into consideration by the model. The Bradley-Terry model was fitted using the 'BTm' function available in the BradleyTerry2 package, that is, the ability scores were estimated using the maximum likelihood function (Varin et al., 2016). In this process, one of the journals is usually set as the 'reference'. The ability score assigned to the reference journal serves as a reference point for interpreting the abilities of all other items in the set. Liner and Amin (2004) note that any journal in the list can be used as the reference journal, and the analysis will yield the same results regardless of which one is chosen. Turner and Firth (2012) have noted that there may be some concerns regarding the Bradley-Terry model's handling of self-citations and the assumption of independence between citations in academic articles. Nevertheless, they mention that several arguments were given by Stigler (1994) in support of using the model despite the concerns raised about its limitations.

In the Bradley-Terry model, the data is typically represented as a directed graph where each item is represented as a node and each pairwise comparison is represented as a directed edge between the two nodes. If the graph is disconnected, meaning that there are two or more separate components that are not connected by any edges, the model may not be able to estimate the abilities of the items accurately. For regularisation, we introduced a dummy journal which equally cites and is cited by all the other journals in the set. The estimated ability scores can be extracted using the function 'BTabilities' of the package. The journals can then be ranked based on the estimated ability scores. The higher the score, the higher the importance of the journal.

The rankings of journals based on citation-based metrics often fail to take uncertainty into account. Measuring the uncertainty associated with rankings helps to provide a more accurate representation of the ranking and its potential variability. In order to account for uncertainty in the journal ranking obtained, we use quasi-variances (Firth & Menezes, 2004). In other words, quasi-variances are used to estimate the uncertainty associated with the ability (export) scores of the journals. Whenever there are a large number of variables, as in this case, a large number of journals, it becomes computationally challenging to calculate the entire covariance matrix. In order to reduce the dimensionality of the covariance matrix, we can use quasi-variances, which are a simplified representation/approximation of the covariance matrix. The quasi-variances are estimated with the help of the ‘qvcalc’ package (Firth, 2020). Finally, the comparison intervals are calculated with the help of these values. To create a 95% comparison interval around the parameter estimate (μ), a basic calculation of (μi±(1.96 × quasi-standard error of each parameter estimate (μi))) is employed (Gayle & Lambert, 2007), where the quasi-standard is the square root of the quasi-variance obtained.

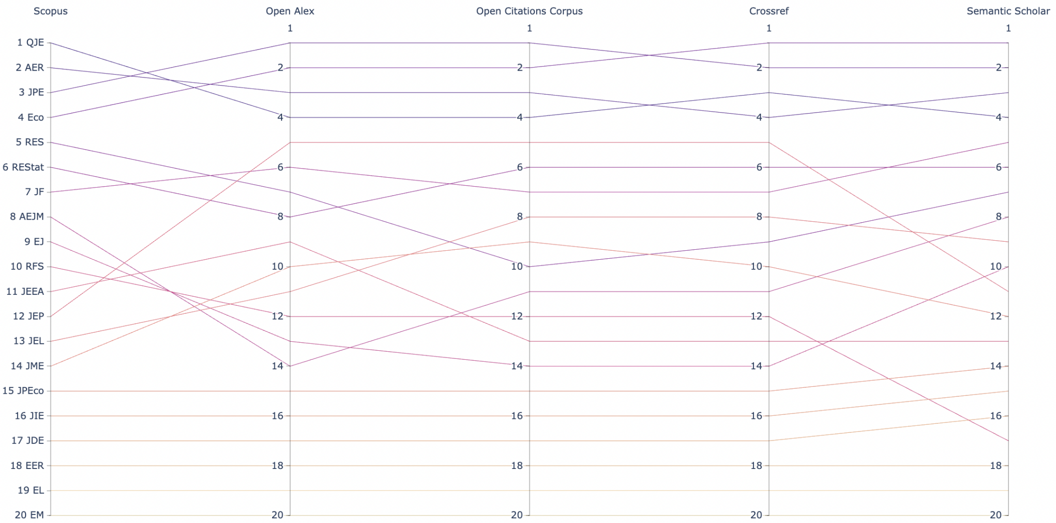
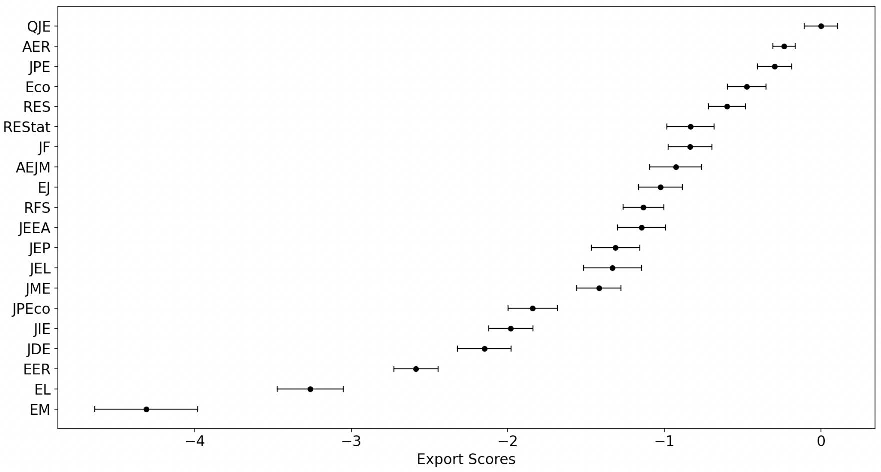


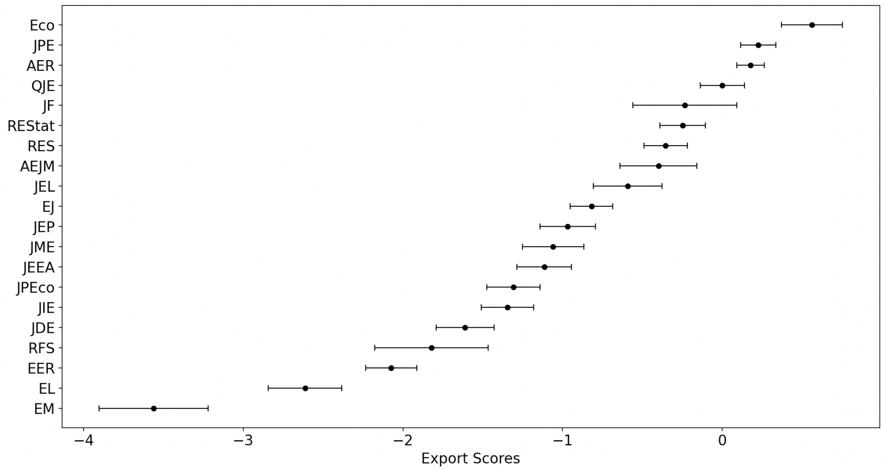
Figure 7.1.: *Ranking of top 20 Economic Journals in 2016-2020 according to the Stigler Model*

7.2. Visualisation and Interpretation of Results

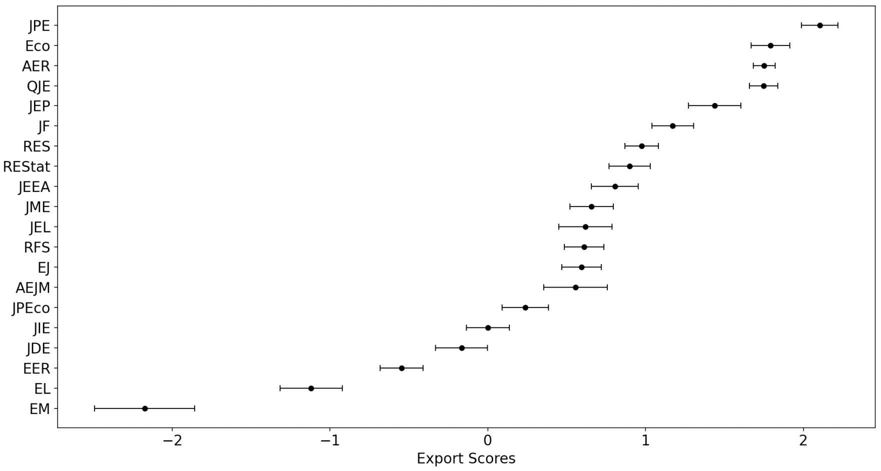
In this section, we interpret the rankings based on the Stigler model. Figure 7.1 depicts the ranking of economic journals obtained using the model. The top 5 journals: QJE is ranked between 1 & 4, AER is ranked between 2 & 4, JPE is ranked between 1 & 3, Eco is ranked between 1 & 4 and RES between 5 & 10. It is encouraging to see that four out of five top economic journals, namely, AER, QJE, JPE, and Eco, consistently rank in the top four among all the five data sources. The fifth journal, RES, consistently ranks in the top 10. Across 4 data sources, the journals JPEco, JIE, JDE, EER, EM and EL have the same ranking, and these journals hold the bottom six positions. However, in the data obtained from Semantic Scholar, there is a slight variation in ranking, and RFS occupies the 17th spot. The rankings for the other nine journals vary significantly depending on the data source. It is observed that, based on OpenAlex and OCC data, the rankings of 10 journals namely JPE, Eco, QJE, AER, JPEco, JIE, JDE, EER, EL and EM, are identical. Additionally, the ranking obtained using the data from OCC and Crossref databases have very similar rankings with slight variations in ranks for some of the journals. This might be due to the fact that Crossref is a major source of the OCC data.



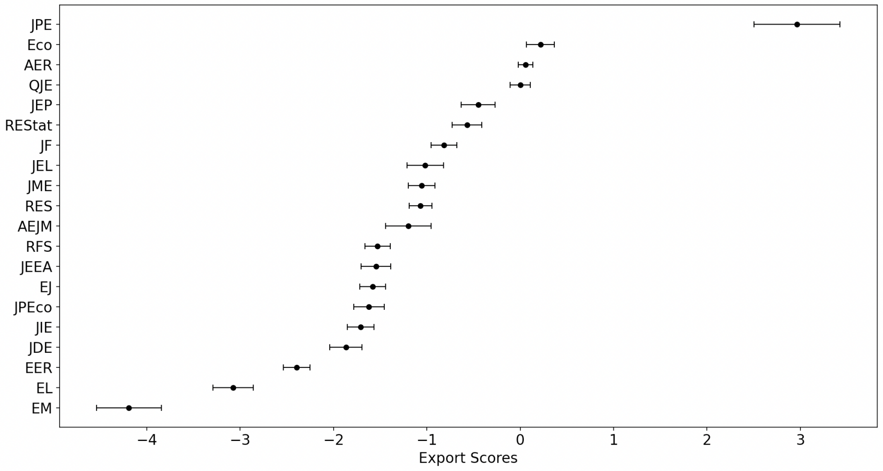
(a) Centipede plot of estimated journal export scores based on Scopus data



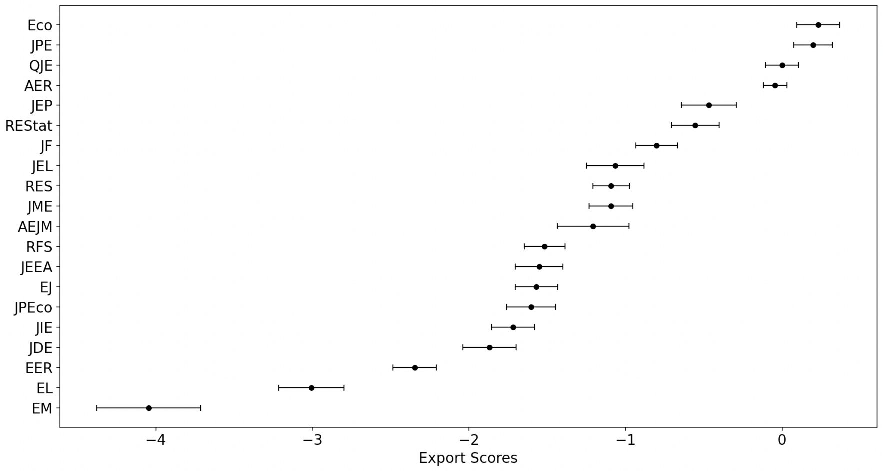
(b) Centipede plot of estimated journal export scores based on Semantic Scholar data



(c) Centipede plot of estimated journal export scores based on OpenAlex data



(d) Centipede plot of estimated journal export scores based on OCC data



(e) Centipede plot of estimated journal export scores based on Crossref data

Figure 7.1.: *Centipede plots illustrating the uncertainty associated with the journal rankings*

Evaluation of results is an integral part of any academic work, and it is essential to have a well-defined success criterion for proper evaluation. Here, the success criterion was to verify that the top five journals in the ranking corresponded with the general consensus about the top five journals in economics. It appears that the results are mostly in line with what was expected. As mentioned earlier, the journals QJE, AER, JPE and Eco occupy the top 4 positions in the ranking obtained and show no deviation from the general consensus. However, the rank of the journal RES is not among the top five and therefore is not in agreement with the expected result. This anomaly could be caused by several reasons, and section 7.4 discusses some of the possible causes.

The level of uncertainty associated with each of the five journal rankings will now be examined separately. Figure 7.1 depicts the uncertainty associated with the rankings using centipede plots. It illustrates the estimated export score of the Stigler model along with the 95% comparison intervals, and the limits are μi±1.96×QSE, where ‘QSE’ denotes the quasi-standard error of the corresponding export score (Varin et al., 2016).

In the Scopus-based ranking, only the comparison intervals of the first journal (QJE), the group of journals ranking 15, 16 and 17 (JPEco, JIE, and JDE), and the last three journals (EER, EL, and EM) are well-separated. The ranking based on OpenAlex data shows a similar behaviour, with the first journal (JPE), the group of journals ranking 2, 3, and 4 (Eco, AER, and QJE) and the last three journals (EER, EL, and EM) having well-separated comparison intervals. The plot based on OCC data exhibits a behaviour similar to that based on OpenAlex data. However, the outstanding position of the journal ranking first (JPE) is easily visible. However, for the ranking based on Semantic Scholar data, the comparison interval of only the first journal (Eco) and the last two journals (EL and EM) are well separated. On the other hand, for Crossref data, the group of the first four journals (Eco, JPE, QJE, AER) and the last three journals (EER, EL and EM) are well separated. It is evident from the centipede plot that most estimates of journal export scores are marked by substantial uncertainty. For most of the journals, with the exception of a few top-ranked and bottom-ranked journals, the estimated export scores differ only by a small amount, but these differences are not statistically significant. The analysis was inspired by findings from Varin et al. (2016), who analysed statistical journals in a similar fashion.

Overall, in the centipede plot based on the data from three of the five data sources (OpenAlex, OCC and Crossref), it can be observed that the ranks of the top four journals as well as the bottom three journals are well estimated. On the other hand, in the centipede plot based on the data from Scopus and Semantic Scholar, only the top-ranked journal and a few lower-ranked journals have well-separated comparison intervals. However, in both cases, many mid-ranking journals have a high degree of uncertainty associated with their ranking as comparison intervals overlap.

In all of the cases, the plausible range of ranks for the majority of journals, excluding those at the very top or bottom, exhibit significant overlap. Accordingly, it can be concluded that the journal rankings are predominantly dependent on data sources, and the rankings, especially the middle-ranked journals, are subject to a high degree of uncertainty.

7.3. Comparison of rankings based on JIF, AIS and Stigler Model

There are many methods/metrics available for ranking journals. Although there have been many studies examining the rankings of economic journals, only very few studies have taken into account the uncertainty associated with them (Lyhagen & Ahlgren, 2020). Popular metrics such as the IF and h-index fail to capture the uncertainty involved. The IF, for instance, is still widely used to evaluate works and individuals (Stephan et al., 2017). Often, a lot of recognition is given to works and journals with high IFs. However, the metric does not capture the uncertainty associated with journal rankings and is based solely on the average number of citations received per article published in the journal during a specific period of time. An attempt to quantify the uncertainty associated with JIFs of economic journals was made by Stern (2013).

A ranking of the selected economics journals based on the JIF, AIS, and the Stigler model is shown in Table 7.1. The metric's abbreviation, along with the data source used for ranking, is indicated in the column header of the table. In a similar study, Varin et al. (2016) compared statistics journals using a much larger number of journals and additional metrics.

Table 7.1.: *Comparison of journal ranks based on JIF, AIS and Stigler Model (SM)*

Rank	JIF (WoS)	AIS (WoS)	SM (Scopus)	SM (OpenAlex)	SM (OCC)	SM (Crossref)	SM (Semantic Scholar)
1	QJE	QJE	QJE	JPE	JPE	Eco	Eco
2	AER	JPE	AER	Eco	Eco	JPE	JPE
3	JPE	AER	JPE	AER	AER	QJE	AER
4	JEL	JF	Eco	QJE	QJE	AER	QJE
5	JEP	Eco	RES	JEP	JEP	JEP	JF
6	JF	RES	REStat	JF	REStat	REStat	REStat
7	REStat	JEP	JF	RES	JF	JF	RES
8	RES	JEL	AEJM	REStat	JEL	JEL	AEJM
9	Eco	RFS	EJ	JEEA	JME	RES	JEL
10	RFS	AEJM	RFS	JME	RES	JME	EJ
11	AEJM	REStat	JEEA	JEL	AEJM	AEJM	JEP
12	JEEA	JEEA	JEP	RFS	RFS	RFS	JME
13	JME	JME	JEL	EJ	JEEA	JEEA	JEEA
14	JDE	EJ	JME	AEJM	EJ	EJ	JPEco
15	JIE	JDE	JPEco	JPEco	JPEco	JPEco	JIE
16	EJ	JPEco	JIE	JIE	JIE	JIE	JDE
17	EM	JIE	JDE	JDE	JDE	JDE	RFS
18	JPEco	EER	EER	EER	EER	EER	EER
19	EER	EL	EL	EL	EL	EL	EL
20	EL	EM	EM	EM	EM	EM	EM

It is evident that these metrics do not provide a common, unambiguous picture of the ranks of these journals. According to Varin et al. (2016), rankings of acceptable quality should place the most prestigious journals prominently. However, only three out of the top five journals are ranked highly based on the JIF. In terms of rankings, 'Econometrica' and 'Review of Statistics' do not make the top five. As measured by AIS, all five leading journals rank within the top six, providing a satisfactory ranking. Among the rankings based on five data sources, four out of five top journals are prominently ranked according to the Stigler model. Moreover, the top 4 journals and the journals JDE, JPEco, JIE, EER, EL and EM occupy a similar position in both Stigler Model and AIS-based rankings. Therefore, the ranking based on the AIS is more similar to the one produced by the Stigler model. This is in agreement with the findings of Selby (2020). He observes that there is a strong positive relationship between the log-transformed AIS and the Stigler Model's export score.

The AIS was introduced to overcome the shortcomings of the IF. It is evident from the rankings that AIS and the Stigler model provide a satisfactory ranking compared to the JIF. The JIF is still in use for research evaluation. Stern (2013) observes that it is common for institutions and countries to provide financial bonuses based on the JIFs of journals in which researchers publish their papers. As the IF measures the popularity rather than the impact of scientific work (Setti, 2013), it is not recommended to use it for research evaluation.

7.4. Discussion

As a next step, let us examine the results of the citation analysis and ranking of journals. It is observed that the ranking differs slightly from the expected ranking of the top 5 economics journals. In light of this outcome, our intention is to examine potential factors that may explain the reasons for variations in rankings among different data sources and to investigate why the resulting rankings did not perfectly align with the general consensus:

To begin with, this may be attributed to variations in the data retrieved from different databases. We saw that the distribution of works published in journals greatly varies across different databases. The ranking of economic journals can be influenced by various factors that are reliant on the data sources, including quality of data, coverage of the publications & citation data, and potential biases. The coverage of the data source can influence the ranking, as the selection of the scientific works included in the database can vary widely. Some data sources have comprehensive coverage of publications in a particular field, while others are more selective in their coverage. For example, some databases like WoS have lower coverage in works related to Arts and Humanities (Vera-Baceta et al., 2019).

The quality of data could be another factor. The quality of the data can affect the accuracy and reliability of the citation analysis and ranking. Some data sources may have incomplete or incorrect data, while others may have more reliable data. For instance, the data obtained from Google Scholar may have incorrect citation counts and may not be highly accurate (Waltman, 2016). Similarly, unstructured citation data accounts for 11% of the data in Crossref, and nearly a third (29%) of the publisher-asserted DOIs have not been verified for their accuracy (Tkaczyk, 2019).

Another reason could be that some data sources may favour publications from certain countries, languages, or disciplines, which can impact the final ranking. For example, if the database primarily gives preference to articles published in English, journals that publish articles in other languages may be inadequately represented, specifically, databases like WoS and Scopus have an overrepresentation of articles in the English language (Vera-Baceta et al., 2019). Consequently, the ranking of journals that publish articles in languages such as Chinese could be impacted.

Further, citations may not be a reliable indicator of the quality of a journal. Belter (2015) has presented a number of reasons to support this argument. An author could cite a paper due to various reasons, for instance, to criticise or correct a researcher's work. The existing bibliometric indicators are not capable of acknowledging this diversity since they treat all citations as having the same significance, irrespective of the motive behind the citation. The author also notes that there is a huge difference in citation patterns between fields. An instance of this can be seen in fields like molecular biology, where researchers tend to cite more than the ones in nursing. Critics of citation-based metrics often point out the significant variation in these metrics across different academic disciplines (Selby, 2020). In addition, most of the works tend to gain citations over time. Therefore, bibliometric indicators based on the citation counts favour older papers than the ones published recently.

There is a possibility that the general consensus regarding the ranking of the top 5 journals, that is, the 'ground truth' ranking, may be doubtful or uncertain. While there may be a general consensus among economists and scholars in the field, the ranking of journals is subjective and can vary based on biases, data sources, and the evaluation criteria used. For example, in a study by Bornmann et al. (2017), the top five journals were found to be QJE, Journal of Financial Economics, JEL, JF, and Eco. Moreover, the field of economics is constantly evolving, and new research avenues and areas of focus may emerge that can challenge the traditional ranking of journals.

7.5. Implications

Journal rankings can vary considerably and are influenced by multiple factors, including the metric used for evaluation, the citation window considered, and the data source used (as shown in section 7.3). It is important to note that no single citation data source provides complete coverage of all journals, and the degree of journal coverage can vary significantly across different disciplines (Walters, 2017). Journals from some fields may be poorly represented, and the results of citation analysis may reflect this disparity between fields. Also, no database is entirely devoid of errors & biases. Therefore, the choice of data source can have a significant impact on the end result.

Wall (2009) observes that “nearly all journal rankings in economics use some weighted average of citations to calculate a journal’s impact”. Walters (2017) points out that the use of metrics relying on averages (means) to assess the impact of a journal or individual article can be misleading when the distribution of citations is skewed. Even though new citation indicators have been introduced recently, indicators such as the JIF, which is based on the average number of citations received (per article), are still popular and commonly used for journal evaluation. The IF is often considered the ‘gold standard’ for research evaluation (Kovatcheva, 2022).

Chapter 6 highlighted the fact that the distribution of citations in journal articles is heavily skewed. The study conducted by Stern (2013) aligns with this finding, where he observes that “previous research on the distribution of citations to articles in economics found that the distribution of citations to articles in a journal is skewed”. He also argues that the high IFs of top journals depended mostly on attracting a few highly cited papers. As a result, relying on a single metric, such as the IF, may not present an accurate assessment of journal ranking.

Also, it is important to acknowledge that no ranking system is fully accurate or robust. Despite the availability of a variety of citation-based metrics, there is no single definitive method that can be used to rank journals accurately and reliably. Different ranking systems have different strengths and weaknesses, and they may be more or less appropriate for different purposes and contexts. Therefore, relying on one ranking system or metric may not provide a complete picture. Ranking journals or researchers based on metrics like the IF can further exacerbate the problem since it does not measure the actual impact of a work but rather its popularity (Setti, 2013).

Setti (2013) also remarks that “the scientific impact of journals as evaluated by bibliometrics is a complicated, multi-dimensional construct which cannot be captured by any single measure”. He also observes that using multiple indicators gives a balanced picture of a journal’s impact and reduces the incentive for individuals to manipulate metrics.

Therefore, it is essential to use multiple metrics based on data from different data sources when evaluating the quality and significance of scientific research and journals to get a better perspective of the rankings. Hence, it is helpful to have a tool for extracting and aggregating data from multiple sources, such as the one developed in this dissertation. Furthermore, citation-based metrics can be calculated based on the output obtained, i.e., the cross-citation table, in order to rank journals. Using multiple citation-based metrics based on data from different citation databases can provide a more well-rounded view of an entity's (for e.g., journal's) impact. It can help overcome potential biases and limitations associated with relying on a single database or metric. Therefore, leveraging multiple databases and metrics can lead to a more robust and reliable evaluation of research. It is important to note, however, that citation-based metrics are not without their limitations, and they should be used in conjunction with other methods, such as peer review.

8. Software Description of CitaTrack

We have developed the tool CitaTrack to assist users in extracting citation data. In general, extracting data from citation databases requires a technical understanding of either data mining techniques or REST APIs. Nonetheless, citation data can be extracted using this package even by users without extensive technical expertise. It can be used by researchers, data analysts or bibliometricians alike. However, a basic understanding of Python is beneficial.

8.1. Architecture of the Tool

Figure 8.1 depicts a brief overview of the components of the tool and the data retrieval process. Three key components of the tool are data input, the retrieval and processing of the data using the package, and output. Further information about each component is mentioned in section 8.3. The tool accepts user-generated input files of *.txt* format, performs the retrieval & processing of data, and exports the output file.

The tool is mainly based on the requests package, which allows one to send HTTP requests easily and is one of the most downloaded packages in PyPI today (Reitz, 2022). The main package contains a sub-package for each citation database, each containing a separate class for retrieving and manipulating data. Data can be retrieved by creating instances of these classes and then calling the function `get_data()` for each class. The function initiates the retrieval and processing of data. The parameters for these functions are the date range, the path to the input file and the metadata retrieval option. Using the option, a user can choose between retrieving citation data for a set of journals or a set of authors. Snippets of code are provided in Appendix A.1.

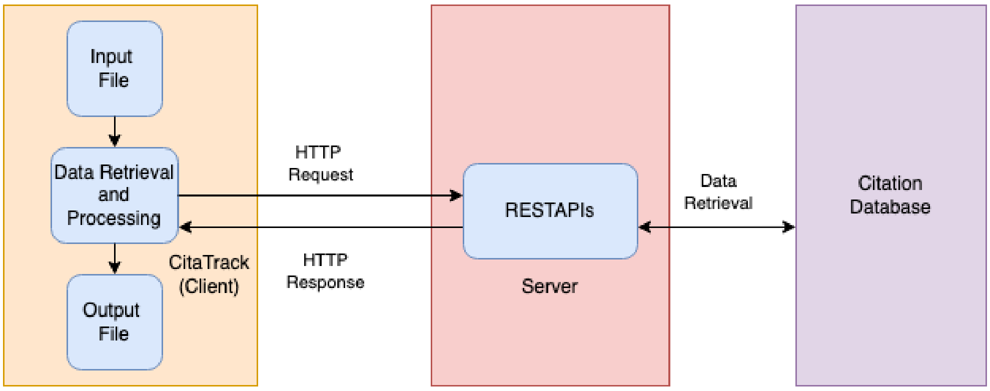


Figure 8.1.: *Block diagram depicting the data retrieval process*

The appropriate API endpoints are then used, and HTTP requests are sent out to retrieve citation data from the respective data sources. Once the data is retrieved, it is processed and cleaned. Once the retrieval and processing of the data are successfully completed, the output is exported as a *CSV* file containing a cross-citation table/matrix of journals/authors.

8.1.1. Cross-Citation Matrix

The distribution of references in a set of journals can be modelled with the help of a cross-citation table or a matrix. For example, a journal-to-journal cross-citation matrix consists of journal titles in both rows and columns. Each cell of the matrix represented by the combination of a row (Citing journal) and a column (Cited journal) of the matrix contains the number of citations received by the cited journal from the citing journal. This concept of journal cross-citation matrices can be extended to authors as well. The uses of cross-citation matrices have been mentioned briefly in Chapter 1 and 2.

As observed by Todorov and Glänzel (1988), the number of citations received by a specified set of journals (in the matrix) can be used to establish a simple (row) ranking of journals. A citation matrix can be transformed into a cross-citation table, where each cell of the matrix, specifically the citation counts, can be represented as a pair of citing and cited journals as depicted in Fig 8.1. The table of cross-citations provides data in a suitable format for post-processing. Also, the logical arrangement of data makes it easy to understand and use.

8.2. Citation Databases Used

For the development of the tool, OpenAlex, Scopus, Crossref, OCC, and Semantic Scholar, in conjunction with Crossref, were used as data sources. Details of the data retrieval process of each database and the APIs used are discussed in the upcoming section.

Previously, the field of bibliometrics was governed by the data retrieved from proprietary sources. However, open availability of the citation data is critical to the field because it "is essential to promote reproducibility and appraisal of research, reduce misconduct, and ensure equitable access to and participation in science" (Sugimoto et al., 2017). Among the citation databases used as data sources for the development of the tool, all the databases except Scopus provide open access. In addition, all of the data sources are multidisciplinary. It is worth noting that multidisciplinary citation databases are fewer in number among all the citation databases available.

The metadata retrieved by the tool from the citation databases is related to the three most important entities namely, works, authors and sources.
Work: Works are scholarly documents like articles, books, datasets and theses.

Author: Authors are the people who create works.
Source: Sources are where the works are hosted.

Among the data extraction methods mentioned in section 3.1, APIs were used to retrieve citation data, and wrappers were built around each of the APIs used. The database snapshot was not used due to the additional complexity of mining big data. Moreover, additional tools and resources are needed for data retrieval. Also, some of the organisations, for example - Crossref, do not provide free access to their database snapshot. Hence, the most efficient method for retrieving citation data was to use the APIs. Each data source along with the API endpoints used for the development of CitaTrack, is mentioned in detail in the upcoming sections.

8.2.1. OpenAlex

As mentioned in section 3.2.5, the OpenAlex database is free and an open-source database. There is no authentication required to use the OpenAlex API endpoints. However, they provide two API pools similar to Crossref - the polite pool and the common pool. The polite pool is available to users who identify themselves and provide their email addresses (via the request header). The polite pool has the advantage of having a faster and more consistent response time. The common pool is used by users who do not want to identify themselves and want to remain anonymous. However, the response time is slower and less consistent. Also, there are no rate limits for the APIs. There is, however, a burst rate limit of 10 requests per second. So, sending multiple requests at once might result in an error code 429.

Mainly there are three types of entities namely, work, author & sources. One can retrieve details about a single entity or a group of entities and can also filter & search a list of entities or count/group these entities.

Retrieval of citation data related to a set of journals & authors from the OpenAlex

For the creation of a journal-to-journal or author-to-author cross-citation table, separate endpoints are available to retrieve citation data related to authors or journals. The first step in citation data retrieval is to search for the journal or author by name or ID using the APIs provided by OpenAlex. Below are the details regarding the APIs.

'/source' and '/author' endpoints

These endpoints are used for retrieving the metadata related to the sources and authors of the scientific works. It is possible to filter the data based on specific criteria like display names, Journal IDs, or Author IDs. Since every journal and author has an OpenAlex ID associated with them in the citation database, disambiguation of the names of those entities can be done using those IDs. For instance, to search for a journal using its display name or title, we could use the following example:

Search by Source Name:
https://api.openalex.org/sources?filter=display_name.search:chemistry
&per-page=3

The above request would filter out all the journals containing the word 'chemistry' in their display name. The above query retrieves details of 3 journals having chemistry in their display name. Details about the journals, for instance, ISSN, OpenAlex ID, publisher, cited by count etc., were included in the metadata retrieved. The display names of the three journals retrieved were:

1 Journal of Biological Chemistry
2 Analytical Chemistry
3 Inorganic Chemistry

Similarly, we could use the filter to search for a journal or an author using their OpenAlex ID. An OpenAlex ID is a unique and non-nullable identifier consisting of alphanumeric values given to an entity. It is easy to retrieve the OpenAlex IDs of an entity by looking up the OpenAlex website. Given below is an example of searching different authors in the database using an OpenAlex ID.

Search by Author ID:
https://api.openalex.org/authors?filter=openalex_id:A2761537998
The above HTTP request fetches details regarding an author having a specific OpenAlex ID. Here, for instance, the information retrieved is regarding the author 'Liz E. Tobler-Gómez' and includes other data like ids, last known institution, cited by count etc.

'/works' endpoint

Once we have the metadata related to a particular journal or author, the works published in the journal or written by the author need to be retrieved. The '/works' endpoint can be used to fetch data related to a particular work or a set of works published in a journal or written by a particular author. The journal's or author's ID (from the metadata retrieved previously) can be used as a parameter in the request sent. By doing so, ambiguity caused by similar names or titles is avoided. The following example can be used for retrieving the author's works given the Author's OpenAlex ID

Retrieve the works of a specific author
https://api.openalex.org/works?filter=author.id:A2761537998&per-page=3

The three works of the author 'Christian R. Mejia' can be retrieved using the above request. Details regarding the works, including authorships, host source, referenced works, ids, information regarding open access etc., is part of the metadata retrieved. The titles of the five works that were retrieved are given below:

1 The Peru Approach against the COVID-19 Infodemic: Insights and Strategies
2 Self-medication practices during the COVID-19 pandemic among the adult population in Peru: A cross-sectional survey
3 Rate of gestational weight gain, pre-pregnancy body mass index and preterm birth subtypes: a retrospective cohort study from Peru

Similarly, works published in a specific journal can be retrieved using the filter 'journal.id' and then providing the corresponding OpenAlex ID of the journal. The retrieved metadata contains information regarding the works published in the journal, along with the references of each work. By aggregating and summarising the citations received, cross-citation tables can easily be created once the works and references of each work are available. Similarly, cross-citation tables for other data sources can be generated by using the respective APIs and adapting the data retrieval process slightly. The snippet of code for the construction of the cross-citation table is provided in Appendix A.2.

8.2.2. Scopus

Scopus is one of the subscription-based data sources. As noted in section 3.2.3, Elsevier's developer portal allows one to create an API key, and this key permits one to send HTTP requests via their APIs. One can fetch metadata related to journals, authors, other scientific publications etc., from the database using APIs such as Scopus search. An institutional key can be requested by emailing Elsevier's support team if one is affiliated with an institution that subscribes to Scopus. There are around ten different Scopus APIs available, and each API has a different weekly request quota and rate limit. By specifying the desired format in the HTTP header of the request, the APIs can provide results in either XML or JSON format. Some of the APIs, for example, the 'citations_overview' API, have restricted access and requesting access to the API requires contacting Elsevier's 'Data for Research & Discovery' support team and providing the appropriate reason for using it.

Serial Title API

Serial title API retrieves metadata related to serial titles, in particular, the published sources in the database. Either the display name of the journal to be searched can be specified, or the ISSN of the journal can be used. The API has a weekly quota of 20000 requests and a rate limit of 3 requests per second. One such example request is given below:

Search by journal name
https://api.elsevier.com/content/serial/title?title=
Journal of the Royal Statistical Society. Series B: Statistical Methodology

Author Search

Scopus Search API provides the ability to search authors based on their first and last names or their Scopus Author ID. For example:

Search by Author ID
https://api.elsevier.com/content/search/author?query=authlast(Smith)
%20and%20authfirst(Albert)

The above request returns all the authors having the first name Albert and last name Smith. The search returned details about 96 authors having similar names.

Scopus Search API

Once the existence of the particular journal/author is confirmed in the database, Scopus search API can be used for retrieving information regarding works in a journal. The API is used for retrieving metadata regarding abstracts of works stored in the data source. With this API, one can retrieve metadata related to a journal/author by searching for the display name of the journal or the name of the author. Alternatively, one could also search the database using ISSN or AU-ID. ISSN is an international standard for serial publications, and AU-ID is the author ID for each author in the Scopus database. Author ID is automatically generated if the author has a paper indexed in Elsevier's database (Stevens, 2022). The search API has a weekly quota of 20000 requests and a rate limit of 6 requests per second. As highlighted above, one can retrieve the data regarding works in a journal using the ISSN of the journal. An example request is provided below:

Search by ISSN:
https://api.elsevier.com/content/search/scopus?query=ISSN(0090-5364)
&count=3

This query retrieves three publications from the Annals of Statistics, whose ISSN was supplied as a parameter in the query. However, no information about references of each work or 'cited by' information is available in the response. The citation information of each work has to be fetched separately.

The information about works that cited a particular work can be retrieved via the Scopus search API by using the 'REFEID' function. It records the EID of each article that cites the current one (Selby, 2020). An EID is a unique identifier given to academic works in the Scopus database. The REFEID is a restricted function, and a separate request must be submitted to use it. Access to this function can be requested by clients with an Elsevier subscription. An example request is given below. The EID of the work should be included in the request sent, and the metadata related to other works which have cited the current work can be retrieved as the response from the API.

To retrieve the ‘Cited by’ data:
https://api.elsevier.com/content/search/scopus?query=
REFEID(2-s2.0-85090683915)

Alternatively, the citation information can be retrieved by the ‘Citation Overview API’ published by Scopus. However, this approach was not tested. The citation information fetched from both Citation Overview API and REFEID function only retrieve metadata on the article level. As confirmed by Katherine Ruth Garcia, one of the support executives of Elsevier, currently, there exists no means to retrieve citation information on a journal level. The major disadvantage of aggregating citation data for an entire journal is that each work’s citation information must be retrieved individually, making data retrieval very slow. For journals containing a large number of articles, it is a major disadvantage when analysing citations. The limited number of requests is also another major drawback.

8.2.3. Crossref

Crossref APIs are publicly available and can be accessed to retrieve the needed metadata. No sign-up is needed to use the API, and the data is not subject to any copyright. However, some abstracts contained in the metadata might be subject to copyright by the publishers (Kemp, 2020). There are three API pools - public, polite and plus. One can use Crossref anonymously using the public pool or provide an email address while sending the request to use the polite pool. Polite pool responses are faster and more consistent than public pool responses. For services that require high predictability in terms of traffic patterns, plus service is recommended. With the plus service, one should authenticate oneself using an API key.

The existence of the journal is confirmed using its name or ID via the journal API. Once the needed metadata is fetched, the works in the journal can be retrieved using an additional parameter, ‘/works’, in the path of the same API. The metadata returned contains the bibliographic reference of each work in the journal. The details regarding the journal API are provided below:

Journal API

Journal API can be used to retrieve the works in a journal. A simple query can be used to find journals by their display names. The following request is used to search the database and retrieve journals having the keyword ‘Biometrika’ in the metadata.

Search by Journal name:
https://api.Crossref.org/journals?query=Biometrika

If we have the journal ID, in particular the ISSN, the database can be queried to retrieve the metadata. Additionally, the works published in a journal can be retrieved using the 'works' endpoint:

Search by Journal ID:
https://api.Crossref.org/journals/0033-5533/works?rows=3

Here, '003-5333' is the ISSN of the Quarterly Journal of Economics. The above request can be used to retrieve three works published in the Quarterly Journal of Economics.

Author Data Retrieval

One could use the query.author parameter to retrieve the works by a particular author, as illustrated below. However, the major problem is that when multiple authors have the same name, it is difficult to distinguish between them.

Search by Author's Name
https://api.Crossref.org/works?query.author=Josiah Carberry

The above request returns all works published by the authors with the name "Josiah Carberry". The use of an ID, such as ORCID, can resolve the ambiguity in author names. Currently, Crossref doesn't offer an API for retrieving author details explicitly. As a result, only a search option that uses ORCIDs is offered to eliminate ambiguity in names. Given below is an example of a query which can be used to retrieve works by a particular author.

Search by ORCID:
https://api.Crossref.org/works?filter=orcid:0000-0001-8255-3853

8.2.4. Semantic Scholar

The APIs [1] provided by Semantic Scholar is free to use. Requests can be sent up to 100 per 5 minutes using the API endpoints. For higher rates, an API key must be obtained after filling out an application form on their website. A rate limit of 100 requests per second was requested to perform the citation analysis mentioned in Chapter 7. Two of the major use cases listed are paper and author lookup. One of the major drawbacks of Semantic Scholar is that currently, there is no way to retrieve journal-level information. Currently, only metadata regarding individual papers can be retrieved from the data source. A paper identifier such as DOI or PubMed ID needs to be provided in the request URL to retrieve the works.

[1] https://www.semanticscholar.org/product/api#Documentation

The paper identifiers of individual papers in journals must be obtained first to retrieve the citation information of journals. The DOI was selected as the paper identifier (as Crossref is already being used as a data source). Since DOIs are owned exclusively by Crossref, they can be obtained from the Crossref database. It is possible that Semantic Scholar does not contain all of the Crossref works. Due to time constraints, we have only tested the approach mentioned above and have not explored any other approaches.

Therefore, to aggregate citation counts across journals, the DOIs of each work within each journal are retrieved via Crossref's Journal API. These DOIs can then be used to fetch references for each work using the Paper Lookup API from Semantic Scholar.

Paper Lookup API

A paper can be searched in the database using the API by providing an identifier in the path of the URL. The identifier used for paper lookup can be S2 Paper ID, DOI, ArXiv ID, ACL ID, PubMed ID or Corpus ID. One of the major advantages of using the endpoint is that there is a provision to filter out just the needed metadata. An example request is provided below.

Fetch references of works
https://api.semanticscholar.org/graph/v1/paper/10.1016/J.JOI.2016.02.007?
fields=title,references.title,referenceCount

In the request above, the DOI of a paper in the journal of infometrics is provided as the query parameter. The response retrieved includes just the title of the work, the title of the references of each work and the reference count. This saves resources that would otherwise be required to handle additional metadata.

Author Lookup API

Metadata regarding an author can be retrieved using the author lookup API. An author's first and last name is specified as the query parameter. The response contains all authors having the name provided as the query parameter along with their IDs.

Search by Author Name
https://api.semanticscholar.org/graph/v1/author/search?query
=rachel+adams

It is also possible to search for an author using their author ID. The below request can be used to retrieve the title and publication of all the works published by an author having the ID 145612610.

Search by Author ID
https://api.semanticscholar.org/graph/v1/author/145612610?fields
=papers,papers.publicationDate

8.2.5. OpenCitations corpus in conjunction with Crossref

Using the SPARQL endpoints and REST APIs, all the metadata in the OCC can be accessed. Many users might not be familiar with semantic web technologies, which is why the APIs are provided in addition to SPARQL endpoints. One can authenticate themselves by using an OpenCitations access token. Users can get an API key by inputting their email address on the OCC website[2]. The access token can then be passed in the ‘authorisation’ header while sending the request.

Similar to Semantic Scholar, the OCC database also does not provide a way to retrieve metadata on a journal-level using the APIs. Therefore, it needs to be used in conjunction with Crossref. First, the DOIs of each work in the journal should be retrieved from Crossref. Since the source of the OpenCitations Data is Crossref, most of the data present in Crossref will be present in OCC database as well. The major disadvantage of the process is that it is slow and expensive to query the database to retrieve the citation count of all the works in the journal.

API to fetch references

The bibliographic references can be retrieved via this endpoint after providing the corresponding DOI of the work in the URL. An example of the request URL is given below:

To fetch references of works
https://OpenCitations.net/index/coci/api/v1/references/
10.1515/libr.1996.46.3.149

In order to fetch citation information about the works of authors, the works can first be retrieved using their ORCIDs from Crossref. Then the references of each work can be fetched using the above API.

8.2.6. Additional Parameters

Additional parameters like the publishing year, paging technique, number of items to be retrieved, the format of the response etc., can be specified while sending each HTTP request. These additional parameters help us in filtering out the response returned. In CitaTrack, all the HTTP requests to be sent has a date range as a filter. It is to ensure that only the articles published in

[2] https://opencitations.net/accesstoken

the specific date range are fetched. Additionally, the paging technique used was cursor paging, and the response returned is in JSON format.

8.3. Usage and Implementation

The package is available in the GitLab[3] repository and can be cloned easily. After the package is cloned and installed, it can be imported into a Python Script. Following the import, the user should initialise an instance of the specific citation database they intend to use. Further, the get_data function can be called to retrieve the citation information related to journals or authors.

REST API Calls

REST APIs are used for retrieving the data from the citation data sources. This method of data extraction was preferred over the database snapshots because the latter is a complicated process and requires a lot of resources to extract and mine the database. The OpenAlex website suggests the use of their API endpoints over the database snapshot. These APIs are maintained by the respective institutions maintaining the databases. Different APIs have different authentication and rate limits. So, it helps to use a library for sending the requests to the servers. Requests library in Python is used to make API calls in the tool. The headers containing the authentication information and request parameters should be set before the request is sent. The response received can be further processed.

8.3.1. Input

The input function `get_data()` accepts four parameters, namely: Option, Start Date, End Date and File path. Users should specify a date range as input, and only articles published within the particular date range in the specified journals/by the specified authors will be retrieved. In addition to that, a set of journal names/journal IDs (such as ISSN) or author names/author IDs (such as ORCID) should be input as comma-separated values in a text file. Detailed information about the parameters is as follows:

1. Option: There are two options available for the user. The '1' and '2' options are used to retrieve citation information about a set of authors and journals, respectively.

2. Start date: Date of publication of articles relating to the journal or the author from which the search should be conducted.

3. End date: Date of publication of articles relating to the journal or the author up to which the search should be conducted.

[3] https://gitlab.com/akshayad67/citatrack

4. File path: The absolute path of the text file containing the set of journal or author names.

8.3.2. Data Retrieval and Manipulation

Once the user has provided the necessary information and started the execution of the program, the data retrieval process begins. The first step involves searching the database to verify whether the set of journals or authors specified in the input data is available. An error message is displayed if a journal or author cannot be found. However, the similar names of authors and journals might create some ambiguity. There can also be multiple authors with the same name. Also, abbreviated names such as J. Smith can refer to, e.g., John Smith or James Smith.IDs of the entities are used to resolve this issue. Each entity in a citation database has its own unique ID. For example, OpenAlex uses the OpenAlex ID, Crossref and OCC use the ISSN for journals and ORCID for authors, respectively. An alert message, along with a list of ambiguous journal/author names as well as their IDs, will be displayed on the console. The alert directs the user to search for the author or journal using the ID provided on the console whenever there is ambiguity in the names. Additionally, the ambiguous names along with their IDs, are exported as a CSV file, allowing users to review them in detail. A screenshot of the alert is provided in Appendix A.1.

After the ambiguities are resolved by providing the appropriate IDs, the procedure as given below is followed:

1. The metadata related to all the articles published in a particular journal/by a particular author during the given date range is fetched using the appropriate APIs. This is discussed in detail in section 8.2.

2. Clean the data by extracting the IDs of the articles and list all the references/cited articles contained in those articles if available.

3. If the information about the references/cited articles is not available, an additional step is performed to fetch them from the citation databases using the appropriate APIs.

4. The references/cited articles are filtered according to whether they are published in the desired journals and time period of interest.

5. The number of citations between the journals is then summed up.

8.3.3. Output

The tool aggregates the citation counts between the set of journals or authors (provided as input) into a matrix format. The final output is a CSV file containing a journal-to-journal or author-to-author cross-citation table. An additional CSV file with detailed information about the cited work, citing work,

Table 8.1.: *Cross-citation table for articles published in 2016-2020 in the top five economics journals*

Citing Journal	Cited Journal	Citation Count
The Review of Economic Studies	The American Economic Review	58.0
The Review of Economic Studies	Quarterly Journal of Economics	30.0
The Review of Economic Studies	Econometrica	22.0
The Review of Economic Studies	Journal of Political Economy	47.0
The Review of Economic Studies	The Review of Economic Studies	44.0
The American Economic Review	Quarterly Journal of Economics	59.0
The American Economic Review	The American Economic Review	121.0
The American Economic Review	Journal of Political Economy	53.0
The American Economic Review	The Review of Economic Studies	41.0
The American Economic Review	Econometrica	44.0
Quarterly Journal of Economics	Quarterly Journal of Economics	103.0
Quarterly Journal of Economics	The American Economic Review	83.0
Quarterly Journal of Economics	Journal of Political Economy	45.0
Quarterly Journal of Economics	The Review of Economic Studies	28.0
Quarterly Journal of Economics	Econometrica	41.0
Econometrica	Quarterly Journal of Economics	21.0
Econometrica	The American Economic Review	44.0
Econometrica	Journal of Political Economy	40.0
Econometrica	The Review of Economic Studies	19.0
Econometrica	Econometrica	145.0
Journal of Political Economy	Journal of Political Economy	45.0
Journal of Political Economy	Quarterly Journal of Economics	17.0
Journal of Political Economy	The Review of Economic Studies	16.0
Journal of Political Economy	Econometrica	25.0
Journal of Political Economy	The American Economic Review	6.0

and their journals are also generated. The user can then do a post-processing on this data. Either a data analysis can be done, or the data can be passed as the input to a model, and further insights can be derived. This output can be used as an input to a model or for citation analysis, and further insights can be derived. Fig. 8.1 illustrates the journal-to-journal cross-citation table, while Fig. 8.2 illustrates the author-to-author cross-citation table obtained from OpenAlex data via CitaTrack. In the former, a cross-citation table of the top 5 economics journals for the years 2016-2020 is illustrated, whereas, in the latter, a cross-citation table of a group of authors for the years 2010-2020 is illustrated.

Table 8.2.: *Cross-citation table of a group of authors for the years 2010-2020*

Citing Author	Cited Author	Citation Count
Stephen Reid McLaughlin	Stephen Reid McLaughlin	1.0
Stephen Reid McLaughlin	Jacob G. Levernier	1.0
Stephen Reid McLaughlin	Bastian Greshake Tzovaras	2.0
Stephen Reid McLaughlin	Ariel Rodriguez Romero	1.0
Stephen Reid McLaughlin	Liz E. Tobler-Gómez	1.0
Jacob G. Levernier	Stephen Reid McLaughlin	1.0
Jacob G. Levernier	Jacob G. Levernier	1.0
Jacob G. Levernier	Bastian Greshake Tzovaras	2.0
Jacob G. Levernier	Ariel Rodriguez Romero	1.0
Jacob G. Levernier	Liz E. Tobler-Gómez	1.0
Bastian Greshake Tzovaras	Stephen Reid McLaughlin	1.0
Bastian Greshake Tzovaras	Jacob G. Levernier	1.0
Bastian Greshake Tzovaras	Bastian Greshake Tzovaras	35.0
Bastian Greshake Tzovaras	Ariel Rodriguez Romero	1.0
Bastian Greshake Tzovaras	Liz E. Tobler-Gómez	1.0
Ariel Rodriguez Romero	Stephen Reid McLaughlin	1.0
Ariel Rodriguez Romero	Jacob G. Levernier	1.0
Ariel Rodriguez Romero	Bastian Greshake Tzovaras	2.0
Ariel Rodriguez Romero	Ariel Rodriguez Romero	1.0
Ariel Rodriguez Romero	Liz E. Tobler-Gómez	1.0

8.3.4. User Interface

The tool doesn't have a separate user interface since it's a Python package. In cases where there is an ambiguity in the journal name, author name, or application error, the Python console displays the error. The output (citation data of works in journals) is displayed on the Python console and exported as CSV files.

8.3.5. Best Practices in Software Development

The importance of following best practices cannot be overstated. The development of a tool or software is no exception. Best practices in software development include the use of appropriate coding styles & conventions and creating a design that is easy to use & maintain. Additionally, it is essential to ensure that the code is adequately documented and tested. Following best practices helps ensure that the developed tool is reliable, efficient, user-friendly and functional. Moreover, the code will be easy to maintain and update. The set of best practices followed during the development of the tool CitaTrack is given below:

API Rate Limits

When designing an API wrapper, an important thing to keep in mind is that each API has its own rate limit. Any breach of the rate limit might result in slower access to data and erroneous or no response. This eventuality is taken care of by setting a timeout between the request sent. The value of timeout is set in such a way that the request rate is within the rate limit.

Programming Language

It is important to consider the language the library needs to be written in, the complexity of the tasks it will be used for, compatibility with existing libraries, and how user-friendly it will be. Considering the aspects mentioned, Python is a suitable choice.

Therefore, the tool was developed using the Python programming language, which has numerous advantages over other similar languages such as Java and C++. Python has a comparatively simpler syntax, is more readable and is easy to use. It also has a wide array of libraries that make development easier than in other languages. For instance, it has separate libraries for caching, connecting to the citation database, data retrieval, manipulation and visualisation. The language is considered to be one of the top choices for data analysis, therefore making it suitable for citation analysis as well. The code is entirely written in Python and is organised efficiently using packages and modules. Afterwards, it was packaged and hosted on PyPi.

Caching

Caching is an important step in the data retrieval process of the tool. A cache is a temporary storage which can be used to store data for a limited amount of time. Usually, a user tends to make several similar requests in a short span of time. Hence, caching the responses helps in faster retrieval of data, and reduces the processing time and resource consumption. The ‘requests-cache’ library was used for caching the response, and the SQLite database was used for storing the response data. Despite the fact that there are several Python libraries available for caching, the request-cache library is a better option since both the requests library and the request-cache library are compatible and can be used together.

API Keys

Sending API key as part of URI is fraught with risk and there is a high possibility that it can be compromised. In order to prevent this possibility, the keys are sent as part of the message authorisation header as it is not logged by network elements. This security issue has been taken care of during the development of the tool.

Structure of Codebase

For the codebase to be easily maintained and understood by others, it is essential to design a clear structure and organisation. This was primarily achieved using the OOPs concept of inheritance and an abstract class as the base class. Abstract classes are useful for code reuse, as they provide the basic framework for more specialised classes, making it easier to incorporate new code without having to re-write the existing code. 5 different child classes were created for handling the access to 5 different data sources providing specialised functions needed for the processing of data from each data source. This helps in the modularisation of code and reduction of code design/duplication, as most of the common functionalities are part of the abstract base class. Another advantage of the approach is that it is highly scalable, and it is easy to extend the functionality of the tool to include additional data sources.

Cursor Paging

All the databases offer two types of paginations: first, using rows or offsets and second, cursor paging. The first option can fetch only up to 10k or 20k results depending on the database, and it is slow and expensive compared to cursor paging. Hence, Cursor paging was used as it is more efficient and faster compared to the first option.

Error Handling

Handling errors is critical for the smooth operation of the software. Several exception types have been defined to handle errors efficiently. These exception types have been defined in CitaTrack depending on the type of error that is encountered. Some of the application-specific errors are:

1. Journal not found: If a journal is not found among the set of journals in the user input, this error is thrown.

2. Author not found: This error is thrown when an author is not found among the set of authors specified in the user input.

3. Multiple authors or journals: The error is thrown when one or more journals/authors have similar names.

4. Invalid option: The error is thrown the user inputs an invalid option.

Consistency and Transparency of the Output

The output file generated, that is, the cross-citation matrix after the data retrieval and processing from different data sources is consistent and has a uniform structure & format. To ensure transparency and verifiability of the output, an additional file containing detailed citation information of the entities is parallelly exported as an output.

Readability and Documentation

When developing software, the code should be easy to read and understand, as this will help other developers work with the code more easily. Additionally, ensuring that the code is well-tested is also essential. By doing this, any potential issues can be identified and fixed promptly. Furthermore, ensuring that the code is up to date with any changes to the underlying API is also important, as this can help prevent any potential issues. Finally, thorough documentation of the API is essential to ensure that developers can use the API properly.

9. Conclusion

In this chapter, the focus will be on the limitations of the study, as well as the implications for the academic community and future research directions. Additionally, we have highlighted the major challenges encountered during the construction of the author-to-author matrix. Finally, we conclude the dissertation with a section that provides a summary and key implications of the study.

9.1. Limitations

Limitations are inherent to all studies. The first limitation of this study is related to the number of journals used for ranking. A bigger pool of journals would have provided a more accurate representation of the ranking. However, retrieving citation data can be a time-consuming task, particularly when dealing with different databases, as it is subject to API rate limits or request quotas, which slows down the process.

The approach of ranking journals might not be entirely robust as the cross-citation counts of journals depend heavily on the coverage of citations in the database. Certain bibliographic references may be unavailable in some databases because those works have not been indexed in the databases.

Further, data quality issues in the databases also pose a problem. For example, many of the references are unstructured in Crossref, that is, not all references in the citation data provided by Crossref are structured in a consistent and easily parseable format. Some references may lack key information, such as author names or publication titles, while others may have formatting inconsistencies that make it difficult to match them accurately to their cited sources. These unstructured references pose a challenge for researchers who rely on accurate and complete citation data for their studies. It is difficult to extract the appropriate citation information from these unstructured references in Crossref. Tkaczyk (2019) observed that Crossref has around 11% fully unstructured references. The conversion of unstructured references into structured references is a complex issue that is still being researched.

Due to time constraints, there was a limited investigation into the creation of cross-citation matrices among authors. Extensive studies are required to address several issues related to this topic. Some of the issues are discussed in the next section.

9.1.1. Challenges during the creation of author-to-author cross-citation matrices

There are some challenges during the creation of author-to-author matrices. One of the major challenges is the distribution of citations in papers having multiple authors. For instance, a work may have multiple authors, and when a citation is received for a particular work, it is to be decided how the citation should be shared among the authors. Various approaches have been proposed to deal with this issue, for instance, full and fractional counting (Perianes-Rodriguez et al., 2016). In the former method, for example, if an article written by two authors is cited once, each author would receive one citation, whereas, in the latter, each author would receive half a citation. However, the amount of contribution made by each author to a paper may vary, and many different approaches are available to address this issue. Refer to Waltman (2016) for more information.

Another issue is author name disambiguation. Authors may have the same first name and/or last name. There can be variations in the spelling or presentation of author names across different publications, which can make it difficult to accurately identify and match authors. Name disambiguation mechanisms vary between data sources. Despite numerous attempts to address the issue, it remains largely unresolved (Y. Zhang et al., 2018). Further, co-authors have a tendency to cite each other's works. Mutual citations of co-authors give a distorted representation of the actual situation.

9.2. Future Work

There is potential for the tool to be enhanced by the addition of more data sources. Currently, data extraction from only five data sources is supported. However, the scalable architecture of this tool allows it to be easily expanded. Additionally, citation analysis on a larger pool of journals can be performed while taking into account the API request quotas set by the data sources. Research can be done to figure out how to incorporate unstructured references from Crossref into the citation data. Moreover, the topic of distributing credits of the publication to authors can be studied extensively.

9.3. Implications for the academic community

This dissertation promotes automation of the retrieval of citation data as it has a multitude of benefits over manual retrieval of data. It enables bibliometricians to validate their previous results using the tool. CitaTrack can be downloaded for free, and it also supports various use cases related to citation data retrieval, including citation analysis and ranking of journals. We also intend to promote the use of open citation databases like OpenAlex through our example to inspire others to do the same. Also, the cost involved in using the tool is minimal, and most of the data sources supported by the tool are both free and open. Citation analysis and ranking of journals are very common in fields like economics. Thus, our example highlights the key factors to consider when ranking journals. As an additional motivation, it also demonstrates the importance of using multiple data sources/multiple metrics for ranking journals and evaluating research.

9.4. Conclusion

To sum up, we have covered the fundamentals of citation databases, data extraction, and citation analysis. In response to the challenges associated with these processes, we developed Citatrack. There is a strong need for a tool like Citatrack, or one with similar capabilities, to support the use cases we have discussed. The tool has the potential to greatly benefit the academic community by enabling researchers to conduct reproducible studies with improved accuracy and efficiency. Although Citatrack has effectively addressed some of the challenges associated with citation analysis, there is still potential for improvement that could be achieved through future iterations of the tool.

Despite the challenges and limitations of citation analysis, it remains a valuable tool for evaluating research impact and advancing scholarly communication. Continued efforts to improve bibliometric methods and tools will undoubtedly benefit the research community.

Bibliography

Aksnes, D., & Sivertsen, G. (2019). A criteria-based assessment of the coverage of scopus and web of science. *4*, 1–21. https://doi.org/10.2478/jdis-2019-0001

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *9*. https://doi.org/10.1177/2158244019829575

Aria, M. (2022). A brief introduction to openalexr, accessed: November 2022.

Arum, N. S. (2016). A look at semantic scholar and google scholar comparisons and recommendations.

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *1*, 377–386. https://doi.org/10.1162/qss_a_00019

Bajwa, S. J. S., & Mehdiratta, L. (2021). From traditional bibliometrics to altmetrics: Socialising the research metrics. *65*, 849. https://doi.org/10.4103/ija.ija_1058_21

Bar-Ilan, J., MarkLevene, & Lin, A. (2007). Some measures for comparing citation databases. *1*, 26–34. https://doi.org/10.1016/j.joi.2006.08.001

Barnes, C. (2017). The h-index debate: An introduction for librarians. *43*, 487–494. https://doi.org/10.1016/j.acalib.2017.08.013

Beall, J. (2013). The open-access movement is not really about open access. *11*, 589–597. https://doi.org/10.31269/triplec.v11i2.525

Belter, C. W. (2015). Bibliometric indicators: Opportunities and limits. *103*, 219–221. https://doi.org/10.3163/1536-5050.103.4.014

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *68*, 314–316. https://doi.org/https://doi.org/10.5860/crln.68.5.7804

Besselaar, P. V. D., & Sandström, U. (2019). Measuring researcher independence using bibliometric data: A proposal for a new performance indicator. *14*. https://doi.org/10.1371/journal.pone.0202712

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of science as a data source for research on scientific and scholarly activity. *1*, 363–376. https://doi.org/10.1162/qss_a_00018

Bohannon, J. (2013). Who's afraid of peer review? https://doi.org/https://doi.org/10.1126/science.342.6154.60

Bontis, N., & Serenko, A. (2009). A follow-up ranking of academic journals. *13*, 16–26. https://doi.org/10.1108/13673270910931134

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? an overview of benefits and disadvantages of altmetrics. *8*, 895–903. https://doi.org/10.1016/j.joi.2014.09.005

Bornmann, L., & Williams, R. (2017). Can the journal impact factor be used as a criterion for the selection of junior researchers? a large-scale empirical study based on researcherid data. *11*, 788–799. https://doi.org/10.1016/j.joi.2017.06.001

Bornmann, L., Butz, A., & Wohlrabe, K. (2017). What are the top five journals in economics? a new meta–ranking, 659–675. https://doi.org/10.1080/00036846.2017.1332753

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs i: The method of paired comparisons. *39*, 324–45. https://doi.org/10.2307/2334029

Buchanan, R. A. (2006). Accuracy of cited references: The role of citation databases. *67*, 292–303. https://doi.org/https://doi.org/10.5860/crl.67.4.292

B.V., E. (2022). Elsevier research products apis, october 2022.

Card, D., & DellaVigna, S. (2013). Nine facts about top journals in economics. *51*, 144–161. https://doi.org/10.1257/jel.51.1.144

Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C., & Ram, K. (2021). Rcrossref introduction, accessed: November 2022.

Chang, A.-L., McAleer, M., & Oxley, L. (2010). What makes a great journal great in economics? the singer not the song. *Journal of economic surveys*, *25*, 326–361. https://doi.org/10.1111/j.1467-6419.2010.00648.x

Chudlarský, T., & Dvořák, J. (2017). Can crossref citations replace web of science for research evaluation? the share of open citations. *5*, 35–42. https://doi.org/10.2478/jdis-2020-0037

Clarivate. (2021). Article influence score, accessed: January, 2023.

Clarivate. (2022). The history of isi and the work of eugene garfield, accessed: November, 2022.

Clarivate. (2023). History of citation indexing, accessed: March, 2023.

Colledge, L., James, C., Azoulay, N., Meester, W., & Plume, A. (2017). Citescore metrics are suitable to address different situations – a case study. *European Science Editing*, *43*, 27–31. https://doi.org/10.1016/j.joi.2014.09.005

Collins, S. (2018). *Introducing crossref, the basics* (Technical Report). Crossref. https://doi.org/10.25012/blog.13.11.2018

Cooper, I. D. (2015). Bibliometrics basics. *103*, 217–218. https://doi.org/10.3163/1536-5050.103.4.013

Cox, A., Gadd, E., Petersohn, S., & Sbaffi, L. (2019). Competencies for bibliometrics. *51*, 746–762. https://doi.org/10.1177/0961000617728111

Crossref. (2023). About us,, accessed: November, 2022.

Debackere, K., Verbeek, A., Luwel, M., & Zimmermann, E. (2002). Measuring progress and evolution in science and technology - ii: The multiple uses of technometric indicators. *4*, 213–231. https://doi.org/10.1111/1468-2370.00085

Eysenbach, G. (2006). Citation advantage of open access articles. *4*, e157. https://doi.org/10.1371/journal.pbio.0040157

Firth, D. (2020). *Qvcalc: Quasi variances for factor effects in statistical models, accessed: March 2023.*

Firth, D., & Menezes, R. X. D. (2004). Quasi-variances. *91*, 65–80. https://doi.org/https://doi.org/10.1093/biomet/91.1.65

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). Empirical analysis and classification of database errors in scopus and web of science. *10*, 933–953. https://doi.org/10.1016/j.joi.2016.07.003

Fricke, S. (2018). Semantic scholar. *106*. https://doi.org/10.5195/jmla.2018.280

Garfield, E. (1963). Citation indexes in sociological and historical research. *14*, 289–291. https://doi.org/10.1002/asi.5090140405

Garfield, E. (1965). Can citation indexing be automated? *Statistical Assoc. Methods for Mechanized Documentation, Symposium Proceedings*, *1*.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*. https://doi.org/10.1126/science.178.4060.471

Garfield, E. (1994). The clarivate analytics impact factor, accessed: March 2023.

Gayle, V., & Lambert, P. S. (2007). Using quasi-variance to communicate sociological results from statistical models. *41*, 1191–1208. https://doi.org/10.1177/0038038507084830

Gingras, Y. (2016). *Bibliometrics and research evaluation: Uses and abuses.* https://doi.org/10.7551/mitpress/10719.001.0001

Grech, V., & Rizk, D. E. E. (2018). Increasing importance of research metrics: Journal impact factor and h-index. *International Urogynecology Journal*, *29*, 619–620. https://doi.org/10.1007/s00192-018-3604-8

Guerrero-Botea, V. P., & Moya-Anegón, F. (2012). A further step forward in measuring journals' scientific prestige: The sjr2 indicator. *6*, 674–688. https://doi.org/10.1016/j.joi.2012.07.001

Harnad, S. (2015). Open access: What, where, when, how and why. *Ethics, Science, Technology, and Engineering: An International Resource.*

Harzing, A. W., & Alakangas, S. (2017). Microsoft academic: Is the phoenix getting wings. *Scientometrics*, *110*, 371–383. https://doi.org/10.1007/s11192-016-2185-x

Harzing, A.-W. (2019). Two new kids on the block: How do crossref and dimensions compare with google scholar, microsoft academic, scopus and the web of science? *120*, 341–349. https://doi.org/10.1007/s11192-019-03114-y

Haustein, S., & Larivière, V. (2014). *Chapter 8, the use of bibliometrics for assessing research: Possibilities, limitations and adverse effects.*

Heckman, J. J., & Moktan, S. (2020). Publishing and promotion in economics: The tyranny of the top five. *58*, 419–70. https://doi.org/10.3386/w25093

Heibi, I., Peroni, S., & Shotton, D. (2019). Software review: Coci, the opencitations index of crossref open doi-to-doi citations. *121*, 1213–1228. https://doi.org/10.1007/s11192-019-03217-6

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata, 414–427. https://doi.org/10.1162/qss_a_00022

Herrmannova, D., & Knoth, P. (2016). An analysis of the microsoft academic graph. *22*. https://doi.org/10.1045/september2016-herrmannova

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

Hudson, J. (2013). Ranking journals. *123*, F202–F222. https://doi.org/10.1111/ecoj.12064

Hutchins, B. I. (2021). A tipping point for open citation data, 433–437. https://doi.org/10.1162/qss_c_00138

Jensenius, F. R., Htun, M., Samuels, D. J., Singer, D. A., Lawrence, A., & Chwe, M. (2018). The benefits and pitfalls of google scholar. *51*, 820–824. https://doi.org/10.1017/S104909651800094X

Joshi, M. A. (2014). Bibliometric Indicators for Evaluating the Quality of Scientific Publications. *The Journal of Contemporary Dental Practice*, *15*(2), 258–262. https://doi.org/10.5005/jp-journals-10024-1525

Kalaitzidakis, P., Mamuneas, T. P., & Stengos, T. (2011). An updated ranking of academic journals in economics. *44*, 1525–1538. https://doi.org/10.1111/j.1540-5982.2011.01683.x

Kamińska, A. M. (2017). Plos one – a case study of citation analysis of research papers based on the data in an open citation index (the opencitations corpus), 168–186. https://doi.org/10.5281/zenodo.1066316

Kemp, J. (2020). Rest api, accessed: October 2022.

Ketzler, R., & Zimmermann, K. F. (2012). A citation-analysis of economic research institutes.

Kim, H. J., Jeong, Y. K., & Song, M. (2016). Content- and proximity-based author co-citation analysis using citation sentences. *10*, 954–966. https://doi.org/10.1016/j.joi.2016.07.007

Kleist, N., & Enns, K. (2022). Usgs bibliosearch: A python tool to facilitate searching, cleaning, and compiling of literature citations from across multiple databases, version 1.0.0: U.s. geological survey software release. https://doi.org/https://doi.org/10.5066/P9EW8BO5

Koltun, V., & Hafner, D. (2021). The h-index is no longer an effective correlate of scientific reputation. *16*. https://doi.org/10.1371/journal.pone.0253397

Kovatcheva, P. (2022). Science - postgraduates and research support: Bibliometrics and citations analysis.

Kreiner, G. (2016). The slavery of the h-index—measuring the unmeasurable. *10*. https://doi.org/10.3389/fnhum.2016.00556

Liner, G. H., & Amin, M. (2004). Methods of ranking economics journals. *32*, 140–149. https://doi.org/10.1007/BF02298831

Lisciandra, C. (2022). *Are citation metrics a good thing?, accessed: March, 2023.* http://philsci-archive.pitt.edu/21375/

Lyhagen, J., & Ahlgren, P. (2020). Uncertainty and the ranking of economics journals. *125*, 2545–2560. https://doi.org/10.1007/s11192-020-03681-5

MacRoberts, M., & MacRoberts, B. (1989). Problems of citation analysis: A critical review. *40*, 342–349. https://doi.org/https://doi.org/10.1002/(SICI)1097-4571(198909)40:5⟨342::AID-ASI7⟩3.0.CO;2-U

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & López-Cózar, E. D. (2016). The counting house, measuring those who count: Presence of bibliometrics, scientometrics, informetrics, webometrics and altmetrics in google scholar citations, researcherid, researchgate, mendeley and twitter. https://doi.org/10.13140/RG.2.1.4814.4402/1

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E. D. (2018). Google scholar, web of science, and scopus: A systematic comparison of citations in 252 subject categories. *12*, 1160–1177. https://doi.org/https://doi.org/10.1016/j.joi.2018.09.002

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E. D. (2021). Google scholar, microsoft academic, scopus, dimensions, web of science, and opencitations' coci: A multidisciplinary comparison of coverage via citations. *126*, 871–906. https://doi.org/10.1007/s11192-020-03690-4

Mason, S., & Singh, L. (2022). When a journal is both at the 'top' and the 'bottom': The illogicality of conflating citation-based metrics with quality. *Scientometrics*, 3683–3694. https://doi.org/10.1007/s11192-022-04402-w

McBurney, M., & Novak, P. (2002). What is bibliometrics and why should you care? *Proceedings. IEEE International Professional Communication Conference*, 108–114. https://doi.org/10.1109/IPCC.2002.1049094

McCullough, R. (2022). Citescore 2021 value are now live!, accessed: November, 2022.

McKiernan, E. C., Schimanski, L. A., Nieves, C. M., Matthias, L., Niles, M. T., & Alperin, J. P. (2019). Meta-research: Use of the journal impact factor in academic review, promotion, and tenure evaluations. *8*, e47338. https://doi.org/https://doi.org/10.7554/eLife.47338

McLaren, C. D., & Bruner, M. W. (2022). Citation network analysis. *15*, 179–198. https://doi.org/10.1080/1750984X.2021.1989705

Meho, L. I. (2007). The rise and rise of citation analysis. *20*, 32–36. https://doi.org/10.1088/2058-7058/20/1/33

Mongeon, P., & Paul-Hus, A. (2014). The journal coverage of bibliometric databases: A comparison of scopus and web of science. https://doi.org/10.13140/2.1.4759.7762

Neuhaus, C., & Daniel, H.-D. (2008). Data sources for performing citation analysis: An overview. *Journal of Documentation*, *64*, 193–210. https://doi.org/10.1108/00220410810858010

Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, *41*, 609–641. https://doi.org/10.1002/aris.2007.1440410120

Noorden, R. V. (2016). Controversial impact factor gets heavyweight rival. *540*, 325–326. https://doi.org/https://doi.org/10.1038/nature.2016.21131

OurResearch. (2021). We're building a replacement for microsoft academic graph, accessed: October 2022.

Pajić, D. (2015). On the stability of citation-based journal rankings. *9*, 990–1006. https://doi.org/https://doi.org/10.1016/j.joi.2015.08.005

Perianes-Rodriguez, A., Waltman, L., & van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *10*, 1178–1195. https://doi.org/https://doi.org/10.1016/j.joi.2016.10.006

Peroni, S., & Shotton, D. (2020). Opencitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, *1*, 428–444. https://doi.org/10.1162/qss_a_00023

Peroni, S., Shotton, D., & Vitali, F. (2017). One year of the opencitations corpus. *International Semantic Web Conference*, 184–192. https://doi.org/10.1007/978-3-319-68204-4_19

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of oa: A large-scale analysis of the prevalence and impact of open access articles. *10.7717/peerj.4375*, *6*, e4375.

Piwowar, H., Ferguson, L. M., Schrader, A. C., & Weisweiler, N. L. (2022). Open science factsheet no. 9 based on the 64th online seminar: Openalex. https://doi.org/https://doi.org/10.48440/os.helmholtz.046

Prathap, G. (2017). Eugene garfield: From the metrics of science to the science of metrics. *114*, 637–650. https://doi.org/10.1007/s11192-017-2525-5

Priem, J., Piwowar, H., & Orr, R. (2022). Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. https://doi.org/https://doi.org/10.48550/ARXIV.2205.01833

Pritchard, A. (1969). Statistical bibliography or bibliometrics? *25*, 348–349.

Qiu, J., Zhao, R., Yang, S., & Dong, K. (2017). *Chapter 8, methods of citation analysis*. https://doi.org/10.1007/978-981-10-4032-0_8

Reitz, K. (2022). Requests 2.28.1, accessed: December, 2022.

Retzer, V., & Jurasinski, G. (2009). Towards objectivity in research evaluation using bibliometric indicators – a protocol for incorporating complexity. *10*, 393–400. https://doi.org/10.1016/j.baae.2008.09.001

Ritzberger, K. (2008). A ranking of journals in economics and related fields. *9*, 402–430. https://doi.org/10.1111/j.1468-0475.2008.00447.x

Roldan-Valadez, E., Salazar-Ruiz, S. Y., Ibarra-Contreras, R., & Rios, C. (2019). Current concepts on bibliometrics: A brief review about impact factor, eigenfactor score, citescore, scimago journal rank, source-normalised impact per paper, h-index, and alternative metrics. *188*, 939–951. https://doi.org/10.1007/s11845-018-1936-5

Rose, M. E., & Kitchin, J. R. (2019). Pybliometrics: Scriptable bibliometrics using a python interface to scopus. *10*. https://doi.org/https://doi.org/10.1016/j.softx.2019.100263

Sainaghi, R., Phillips, P., Baggio, R., & Mauri, A. (2018). Cross-citation and authorship analysis of hotel performance studies. *73*, 75–84. https://doi.org/10.1016/j.ijhm.2018.02.004

Saralegui, U. (2021). Opencitingpy, accessed: November 2022.

Scheidsteger, T., & Haunschild, R. (2022). Comparison of metadata with relevance for bibliometrics between microsoft academic graph and openalex until 2020 [Publisher: arXiv Version Number: 1]. https://doi.org/10.48550/ARXIV.2206.14168

Scholar, G. (2022). About google scholar, accessed: November 2022.

Selby, D. A. (2020). *Statistical modelling of citation networks, research influence and journal prestige* (Doctoral dissertation). Department of Statistics, University of Warwick, United Kingdom.

Setti, G. (2013). Bibliometric indicators: Why do we need more than one? *IEEE Access, vol. 1*. https://doi.org/10.1109/ACCESS.2013.2261115

Shotton, D. (2017). Milestone for i4oc – open references at crossref exceed 50%, accessed: October 2022.

Silva, D. (2022). Semanticscholar, accessed: November 2022.

Silva, J. A. T. D., & Dobránszki, J. (2018). Multiple versions of the h-index: Cautionary use for formal academic purposes. *115*, 1107–1113. https://doi.org/10.1007/s11192-018-2680-3

Silva, J. A. T. D., & Memon, A. R. (2017). Citescore: A cite for sore eyes, or a valuable, transparent metric? *111*, 553–556. https://doi.org/10.1007/s11192-017-2250-0

Singh, V. K., Singh, P., Karmakar, M., Leta2, J., & Mayr, P. (2021). The journal coverage of web of science, scopus and dimensions: A comparative analysis. *126*, 5113–42. https://doi.org/https://doi.org/10.1007/s11192-021-03948-5

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. (, & Wang, K. (2015). An overview of microsoft academic service (mas) and applications. *24th International Conference on World Wide Web (WWW '15 Companion*, 243–246. https://doi.org/10.1145/2740908.2742839

Stephan, P., Veugelers, R., & Wang, J. (2017). Reviewers are blinkered by bibliometrics. *544*, 411–412. https://doi.org/10.1038/544411a

Stern, D. I. (2013). Uncertainty measures for economics journal impact factors. *51*, 173–89. https://doi.org/10.1257/jel.51.1.173

Stevens, J. (2022). What is scopus author id?, accessed: October 2022.

Stigler, S. M. (1994). Citation patterns in the journals of statistics and probability. *9*. https://doi.org/10.1214/ss/1177010655

Suber, P. (2012). *Open access*. https://direct.mit.edu/books/book/3754/open-access

Sugimoto, C. R., Waltman, L., Larivière, V., Eck, N. J. V., Boyack., K. W., Wouters, P., & Rijcke, S. D. (2017). Open citations: A letter from the scientometric community to scholarly publishers, , accessed: October 2022.

Surwase, G., Sagar, A., Kademani, B. S., & Bhanumurthy, K. (2011). Co-citation analysis: An overview. *In Beyond Librarianship: Creativity, Innovation and Discovery, BOSLA, National Conference Proceedings*, 16–17.

Tang, J. (2016). Aminer: Toward understanding big scholar data. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 467–467. https://doi.org/10.1145/2835776.2835849

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, C. H. J. (2016). The academic, economic and societal impacts of open access: An evidence-based review. *5*, 632. https://doi.org/10.12688/f1000research.8460.3

Thelwall, M. (2020). The pros and cons of the use of altmetrics in research assessment. *2*. https://doi.org/10.29024/sar.10

Thompson, D. F., & Walker, C. K. (2015). A descriptive and historical review of bibliometrics with applications to medical sciences. *35*. https://doi.org/10.1002/phar.1586

Tkaczyk, D. (2019). What if i told you that bibliographic references can be structured?

Todorov, R., & Glänzel, W. (1988). Journal citation measures: A concise review. *14*. https://doi.org/10.1177/016555158801400106

Turner, H., & Firth, D. (2012). Bradley-terry models in r: The bradleyterry2 package. *48*. https://doi.org/10.18637/jss.v048.i09

Varin, C., Cattelan, M., & Firth, D. (2016). Statistical modelling of citation exchange between statistics journals. *179*, 1–318. https://doi.org/10.48550/arXiv.1312.1794

Vera-Baceta, M.-A., Thelwall, M., & Kousha, K. (2019). Web of science and scopus language coverage. *121*, 1803–1813. https://doi.org/10.1007/s11192-019-03264-z

Wade, A. D. (2022). The semantic scholar academic graph (s2ag). *Companion Proceedings of the Web Conference 2022*, 739. https://doi.org/10.1145/3487553.3527147

Wall, H. J. (2009). Don't get skewed over by journal rankings. *9*. https://doi.org/10.2202/1935-1682.2280

Walters, W. H. (2017). Citation-based journal rankings: Key questions, metrics, and data sources. *5*, 22036–22053. https://doi.org/10.1109/ACCESS.2017.2761400

Waltman, L. (2016). A review of the literature on citation impact indicators. *10*, 365–391. https://doi.org/10.1016/j.joi.2016.02.007

Wan, H., Zhang, Y., Zhang, J., & Tang, J. (2018). Aminer: Search and mining of academic social networks. *1*, 58–76. https://doi.org/10.1162/dint_a_00006

Wei, G. (2018). A bibliometric analysis of the top five economics journals during 2012–2016. *33*, 25–59. https://doi.org/10.1111/joes.12260

Williams, A. E. (2017). Altmetrics: An overview and evaluation. *41*, 311–317. https://doi.org/10.1108/OIR-10-2016-0294

Xia, J., Harmon, J. L., Connolly, K. G., Donnelly, R. M., Anderson, M. R., & Howard, H. A. (2014). Who publishes in "predatory" journals? *66*, 1406–1417. https://doi.org/10.1002/asi.23265

Yuen, J. (2018). Comparison of impact factor, eigenfactor metrics, and scimago journal rank indicator and h-index for neurosurgical and spinal surgical journals. *119*, e328–e337. https://doi.org/10.1016/j.wneu.2018.07.144

Zhang, J. T. J., Li, L. Y. J., & Su, L. Z. Z. (2008). Arnetminer: Extraction and mining of academic social networks. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 990–98. https://doi.org/doi/10.1145/1401890.1402008

Zhang, L., Huang, Y., Cheng, Q., & Lu, W. (2020). Mining author identifiers for pubmed by linking to open bibliographic databases, 209–212. https://doi.org/10.1109/QRS-C51114.2020.00043

Zhang, L., & Glänzel, W. (2004). Journal cross-citation matrices reconsidered. tracing the role of individual journals in the communication network. *Proceedings of WIS 2008*.

Zhang, Y., Zhang, F., Yao, P., & Tang, J. (2018). Name disambiguation in aminer: Clustering, maintenance, and human in the loop. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3219819.3219859

Zhixiong, Z., Liping, K., Xiaolin, Z., & Lin, L. (2013). Establishing arxiv china service group to promote open access movement in china. *1*, 58–76. https://doi.org/10.7536/j.issn.0252-3116.2013.01.009

Zientek, L. R., Werner, J. M., Campuzano, M. V., & Nimon, K. (2018). The use of google scholar for research and research dissemination, 39–46. https://doi.org/https://doi.org/10.1002/nha3.20209

Zijlstra, H., & McCullough, R. (2016). Citescore: A new metric to help you choose the right journal, accessed: November, 2022.

Zupic, I., & Čater, T. (2015). Bibliometric methods in management and organization. *18*, 429–472. https://doi.org/10.1177/1094428114562629

A. Installation and Usage

A.1. Installation

Clone the package from GitLab [1]. The object of the data source one wants to use must be initialized. The `get_data()` function is used to retrieve the data. The user has the choice to retrieve data related to either a set of authors or journals. An input text file containing the names or IDs of a set of journals/authors must be created. The `get_data()` function takes in 3 parameters, namely, the option, the start date, the end date and the path to the input file. An example is provided below:

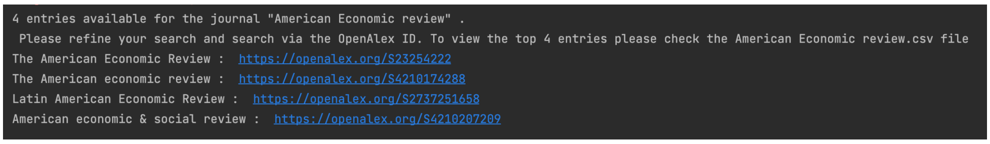
```
from OpenAlex.restful import OpenAlex

#Options 1 and 2
#1 for author
#2 for journal

#Create an object of the data source you want to use
works = OpenAlex()

#Call get_data() to retrieve data
works.get_data(2, ‘‘2018−01−01”, ‘‘2020−12−31”, ‘‘/Users/aa/test.txt”)
```

An example of the input file that contains names and OpenAlex IDs of a set of journals is provided below:

```
https://openalex.org/S141184754,
https://openalex.org/S36178057,
Journal of the European Economic Association,
Journal of Economic Literature,
American Economic Journal: Macroeconomics,
https://openalex.org/S69338747,
https://openalex.org/S23254222,
https://openalex.org/S170137484,
https://openalex.org/S203860005,
https://openalex.org/S95323914,
https://openalex.org/S5353659,
The Review of Economic Studies
```

[1] https://gitlab.com/akshayad67/citatrack

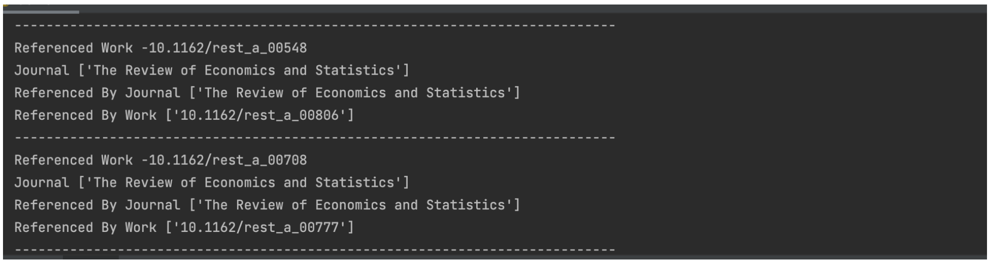
If there are similar or duplicates names in the set, the following error will be displayed.



```
4 entries available for the journal "American Economic review" .
 Please refine your search and search via the OpenAlex ID. To view the top 4 entries please check the American Economic review.csv file
The American Economic Review :  https://openalex.org/S23254222
The American economic review :  https://openalex.org/S4210174288
Latin American Economic Review :  https://openalex.org/S2737251658
American economic & social review :  https://openalex.org/S4210207209
```

Figure A.1.: *Duplicate entry error*

Once the error has been rectified, a cross-citation table will be provided as output and the data will be displayed on the console.



```
----------------------------------------------------------------------------
Referenced Work -10.1162/rest_a_00548
Journal ['The Review of Economics and Statistics']
Referenced By Journal ['The Review of Economics and Statistics']
Referenced By Work ['10.1162/rest_a_00806']
----------------------------------------------------------------------------
Referenced Work -10.1162/rest_a_00708
Journal ['The Review of Economics and Statistics']
Referenced By Journal ['The Review of Economics and Statistics']
Referenced By Work ['10.1162/rest_a_00777']
----------------------------------------------------------------------------
```

Figure A.2.: *An example of the data retrieved as shown on the Python console*

A.2. Code for the creation of cross-citation table

The below snippet of code was utilised for the construction of a cross-citation table. One could adjust the count variable to incorporate the ‘fractional counting’ of citations of the authors.

```
def build_matrix(self, data):
    matrix = defaultdict(lambda: defaultdict(float))
    count = 1
    REF_BY = 'Referenced By '
    for work in data:
        if self.option == 1:
            values = 'Author'
        else:
            values = 'Journal'
        for author in work[values]:
            for referrer in work[REF_BY + values]:
                matrix[referrer][author] += 1.0 / count
```