UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

CLEIBER GUSTAVO SOARES RODRIGUES
JUNIOR

15 Years of Digital RITA - Learning About The Evolution Of The Journal With Bibliometric Analysis

Work presented in partial fulfillment
of the requirements for the degree of
Bachelor in Computer Science

Advisor: Prof. Dr. Marcio Dorn
Coadvisor: Prof. Dr. Manuel Villalobos Cid

Porto Alegre
August 2024

“Os sonhos das pessoas...

Nunca tem fim!”

— MARSHALL D. TEACH

ACKNOWLEDGEMENTS

Escrevo esta seção de agradecimentos em português para que minha família e amigos que não tenham domínio da língua inglesa possam lê-los sem precisar de uma tradução. Muitas pessoas passaram pela minha vida, tanto dentro quanto fora da universidade, e é para estas pessoas que importam a leitura destes agradecimentos.

Primeiramente, gostaria de agradecer a minha mãe que sempre lutou muito com o pouco que teve para me criar da melhor maneira possível e apesar de eventuais divergências sempre me apoiou quando necessário. Mesmo com pouco dinheiro e conhecimento, ela me possibilitou viver o meu objetivo de cursar o ensino superior em uma universidade de excelência. Também vou estender esse trecho para agradecer as minhas irmãs, a quem pude ensinar e aprender muito durante todos esses anos.

Em segundo, gostaria de agradecer a todos os amigos que fiz durante essa longa jornada, pessoas com as quais pude ter o prazer de conhecer, partilhar momentos e dividir sonhos. Conheci pessoas maravilhosas durante todo este trajeto e espero carrega-las comigo para o resto da vida.

Em terceiro lugar, registro aqui meus agradecimentos a todo o corpo técnico docente e administrativo do Instituto de Informática UFRGS. Apesar de todas as adversidades que enfrentei durante minha graduação, conheci muitas pessoas que trabalham arduamente por uma universidade melhor e por um instituto de informática mais plural, e é para estas que vão este agradecimento e meu imenso carinho.

Em quarto lugar, deixo meu agradecimento especial ao meu orientador e co-orientador que me ajudaram muito durante a construção deste trabalho e também me possibilitaram ter um maior contato com diferentes áreas de pesquisa durante esse período.

Em quinto lugar, gostaria de agradecer Zoro, Sanji e Nami, meus gatos, por serem uns amores e ficarem literalmente em cima da mesa me dando suporte emocional.

E por ultimo, gostaria de agradecer ao meu amor, Joana Campos, por toda companhia, carinho e suporte durante essa etapa da minha vida.

ABSTRACT

The Revista de Informática Teórica e Aplicada (RITA) is a scientific journal established in 1989 within the Postgraduate Program of Computer Science at Universidade Federal do Rio Grande do Sul (UFRGS) to promote local publication by producing a high-quality and regular scientific journal. In 2008, the journal reached a new milestone by adopting a new submission and review format after joining Open System Journals, launching its first digital volume, expanding its reach, and attracting new audiences within the scientific community. In 2024, RITA celebrates its 35th anniversary, which is 16 years in digital format, publishing one volume with at least two volumes per year and achieving significant milestones such as indexing by Scopus. Considering the journal's many challenges, this study aims to analyse the materials received over the years to understand better the journal's profile and the contributing factors in this process. We propose a pipeline for conducting bibliometric analysis on the journal's database, allowing us to observe results related to published and declined articles, as well as the authors and institutions that form its network. Lastly, we address some questions based on the results obtained that show us that RITA is indeed on the verge of decline.

Keywords: Bibliometric Analysis. Article Publication. Word Analysis. Data Science. Geographical Visualization.

16 anos de RITA - Aprendendo Sobre A Evolução De Uma Revista Cientifica usando Analise Bibliométrica

RESUMO

A Revista de Informática Teórica e Aplicada é uma revista cientifica criada no Programa de Pós-Graduação em Computação da UFRGS em 1989 com o intuito de fomentar a publicação local, produzindo um periódico científico nacional de alta qualidade e regularidade. Em 2008 a revista atinge um novo patamar de maturidade e passa a estabelecer um novo sistema de submissões e avaliações com a adesão ao Open System Journals e lança sua primeira edição eletrônica, passando a ter mais alcance na comunidade acadêmica. Em 2024 a revista completa 35 anos, sendo 16 destes no digital, mantendo a publicação um volumes com pelo menos dois números por ano, e já conseguiu alcançar conquistas como a indexação pelo Scopus. Tendo em vista que a revista terá muitos desafios pela frente, buscamos neste estudo compreender mais os materiais que recebemos na revista ao longo dos anos de forma a entender o nosso perfil de publicação, o que os autores buscam na revista e os fatores implicantes neste processo. Para isto, propomos um pipeline para aplicar uma analise bibliométrica na base de dados da revista, permitindo analises referentes aos artigos (publicados e rejeitados) e aos autores e instituições que compõem a rede presente aqui. Por fim, abordamos algumas questões com base nos resultados obtidos que nos mostram que a RITA está de fato num periodo de declinio relativo a sua produção.

Palavras-chave: Analise Bibliométrica, Publicação de Artigos, Analise de Palavras, Ciência de Dados, Visualização Geográfica.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AAPP	Author Productivity Per Publication
AGR	Annual Growth Rate
AI	Artificial Intelligence
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CAGR	Cumulative Annual Growth Rate
CC	Collaborative Coefficient
CI	Collaborative Index
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSV	Comma Separated Values
DBLP	Digital Bibliography & Library Project
DC	Degree of Collaboration
DOI	Digital Object Identifier
DT	Doubling Time
EDAS	Editor's Assistant
HTML	HyperText Markup Language
IEEE	Institute of Electrical and Electronics Engineers
IF	Impact Factor
ILS	Library and Information Sciences
MCC	Modified Collaboration Coefficient
NLP	Natural Language Processing
ORCID	Open Researcher and Contributor ID
PDF	Portable Document Format
PNA	Participation of National Authors
PPAA	Productivity Per Author Associated

PPGC	Programa de Pós-Graduação
RGR	Relative Growth Rate
RITA	Revista de Informática Teórica e Aplicada
RQ	Research Question
SBC	Sociedade Brasileira de Computação
SBSEG	Simpósio Brasileiro de Segurança da Informação
SEER	Sistema Eletrônico de Editoração de Revistas
UFRGS	Universidade Federal do Rio Grande do Sul
VAPA	Views Adjusted Per Article
VPA	Views Per Article

CONTENTS

1 INTRODUCTION

In this chapter, we will provide an overview of Revista De Informática Teórica e Aplicada (RITA), its importance for computer science academics in Brazil, the challenges faced in maintaining an Open Journal model, and why it is important to evaluate our production.

1.1 Overview

The Revista de Informática Teórica e Aplicada (RITA) was established in 1989 when public journal access policies weren't popular in Brazil. It aimed to create a free, high-quality, and regularly published national scientific journal dedicated to computer science research, accepting since his conception articles in Portuguese, English, and Spanish (LAMB; OLIVEIRA; GRANVILLE, 2010). The journal became a case of success among university publishers with non-stop production since its launch, reaching national Qualis[1] B and contributes to the democratization of open journal in Brazil and Latin America (LAMB; OLIVEIRA; GRANVILLE, 2007).

In 2008, RITA transitioned to a digital format, significantly advancing its reach and impact. The editorial board moved the project forward by integrating RITA into the Digital Bibliography & Library Project (DBLP) system and utilizing the Sociedade Brasileira de Computação (SBC) Editor's Assistant (EDAS) platform. This shift brought greater attention to the journal, leading to 31 national submissions and ten international ones (OLIVEIRA; LAMB, 2004). After 2010, the journal concluded its physical format and operated exclusively digitally, publishing one volume with at least two numbers yearly. Currently, RITA adopted the Sistema Eletrônico de Editoração de Revistas (SEER) systems provided by Universidade Federal do Rio Grande do Sul (UFRGS) to host his content, ISSN 2175-2745 and can be accessed by the link <seer.ufrgs.br/rita>.

Currently, the RITA editorial team is composed of the editor-in-chief Marcio Dorn with an editorial board formed by 26 members that contribute to the journal publication (<https://seer.ufrgs.br/index.php/rita/about/editorialTeam>), and the editorial policies consists of receiving submissions that haven't been published nor is it before another journal for consideration. Authors should follow the instructions provided by the

[1] Academic journal classification system adopted by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior)

editorial regarding formatting rules, usage of the LaTeX template provided by the journal, and the guidelines related to submission. The journal's privacy statement consists of the decision to publish the paper based only on the paper's content and confidentiality. It also commits to a peer-review process that must be objective and confidential (<https://seer.ufrgs.br/rita/AuthorGuidelines>).

The editorial board has consistently adopted the latest technologies, believing that this strategy would expand RITA's audience and increase submissions in the coming years (OLIVEIRA; LAMB, 2004). Over the past two decades, RITA has achieved new milestones, such as indexing by prominent databases like Google Scholar (1), Latindex (2), Crossref (3), and Scopus (4). These accomplishments suggest that the journal is still progressing in the right direction despite facing numerous challenges.

1.2 Digital publishing trends

Digital publishers have rapidly gained popularity because their articles can reach a wider audience, and the format allows for fast distribution, integrating new research findings quickly and increasing their relevance (OLIVEIRA; LAMB, 2004). Today, there are many free journals, and to maintain their relevance, they must adhere to several key characteristics: specialization, peer-reviewing, transparency, valuing reproducible data, and keeping an author-friendly approach (EDER; FRINGS, 2018). RITA was created in Brazil to achieve these objectives and meets all standards but publishes less frequently than other great publishers like Elsevier, SAGE, and Institute of Electrical and Electronics Engineers (IEEE) (PACHER, 2022).

Another significant challenge for open-access journals is the peer-reviewing process. Finding and retaining qualified peer reviewers is difficult because there is usually no incentive for this work. This lack of incentive increases the waiting time for responses, as peer reviewing is not a top priority for those invited to participate (KUMAR; AHMED, 2022).

1.3 Editorial challenges

Despite the daily challenges faced by open-access journals, in addition to communicating, the main goal of every journal is to achieve better metrics. Improved

indicators, such as the Impact Factor (IF), directly influence the journal perception by external authors seeking a venue to submit their findings (GREENWOOD, 2007). Bridging the gap between a quality and a prestigious journal is challenging, especially since prestigious journals offer authors enhanced academic status (SUBER, 2016). Therefore, it is crucial for the editorial board to closely monitor their publication line to ensure transparency in their achievements and to have solid evidence when action is needed to prevent problems (SAGE PUBLICATIONS, 2023).

1.4 Computer Science Journals in Brazil

When we look at the computer science editorial composition scenario, the volume of journals available is significantly small compared to other countries. The table 1.1 lists the journals focused on computer science created in Brazil with the Qualis classification associated.

Table 1.1: Table of Computer Science journals in Brazil - Evaluation 2017 to 2020

ISSN	*Name*	*Website*	*Qualis*	*Release year*
1678-4804	JOURNAL OF THE BRAZILIAN COMPUTER SOCIETY	<https://journals-sol.sbc.org.br/index.php/jbcs/>	A2	2010
2317-6121	REVISTA BRASILEIRA DE INFORMÁTICA NA EDUCAÇÃO	<https://journals-sol.sbc.org.br/index.php/rbie>	A4	1997
1677-3071	REVISTA ELETRÔNICA DE SISTEMAS DE INFORMAÇÃO	<https://www.periodicosibepes.org.br/index.php/reinfo>	B1	2022
1984-2902	REVISTA BRASILEIRA DE SISTEMAS DE INFORMAÇÃO	<https://seer.unirio.br/index.php/isys/>	B2	2008
2175-2745	REVISTA DE INFORMÁTICA TEÓRICA E APLICADA	<https://seer.ufrgs.br/rita>	B3	1989
2176-6649	REVISTA BRASILEIRA DE COMPUTAÇÃO APLICADA	<https://seer.upf.br/index.php/rbca>	B3	2009
2675-1828	REVISTA BRASILEIRA EM TECNOLOGIA DA INFORMAÇÃO	<https://www.fateccampinas.com.br/rbti/index.php/fatec>	B4	2019
2237-2903	REVISTA DE SISTEMAS E COMPUTAÇÃO	<https://revistas.unifacs.br/index.php/rsc>	B4	2011
1807-4545	INFOCOMP	<https://www.ufla.br/dcom/2005/05/19/ciencia-da-computacao-edita-revista-infocomp/>	B4	1999
1809-5585	REVISTA DE INFORMÁTICA APLICADA	<https://seer.uscs.edu.br/index.php/revista_informatica_aplicada>	C	2005

Source: The Author

The fact that we have a small count of journals can be traced to some factors: the difficulty of maintaining the editorial process due to budget constraints, the lack of incentives, and changes in the evaluation methods established by research funding agencies, which tends to value more researchers who publish in international high impact journals. Another observation that can be taken from this is that despite RITA being the older journal focused on computer science, their classification isn't the better-ranked publication in this scenario, which isn't expected. With that in mind, we want to look more in-depth at RITA production and gain insights that justify the current rank attributed by RITA.

1.5 Work Objectives

Despite RITA existence and importance in distributing national computer science articles in Brazil, we didn't have any form of quantitative studies that give us an overview of the journal to understand its current status. With that in mind, the purpose of this work is to evaluate the production aspects of RITA by using a Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology for extract article information, using this approach to prepare data, calculate metrics, and exploring not only data from published studies but also data from rejected articles to understand why RITA hasn't achieved a better classification despite having more publishing time than the other journals. The Research Questions (RQs) defined to investigate this are defined below:

- RQ1: How has the number of publications and publication trends evolved over the past 15 years, and which years had the highest and lowest number of publications?
- RQ2: How do different bibliometric indices assess the quality and impact of scientific publications in RITA over the last 15 years?
- RQ3: What are the most viewed topics and articles over the past 15 years, and what characteristics do these popular articles share?
- RQ4: Which authors and institutions have the most published and rejected articles, and what is their collaboration network?
- RQ5: What are the key differences between accepted and rejected articles regarding keywords, topics, and other relevant metrics?

These questions are elaborated not only for metrics evaluation but also to identify characteristics that can be useful for the editorial board to achieve better scores and more audience in the academic world. Since our questions want to explore and compare some aspects of both submissions accepted and rejected, some results are anonymized since the objective is to observe properties like spaciality, keyword frequency, and distribution.

The structure of this study is described here: In chapter 2, the present work shows some popular review methods on literature and compares their properties and related works of bibliometric analysis. Chapter 3, a methodology for the experiment is explained in more detail to create a way to execute bibliometrics. Chapter 4 shows the results obtained after executing all steps planned in chapter 3. At the end, chapter 5 has the conclusion around the findings with the branches opened for future works.

2 BACKGROUND AND PREVIOUS WORKS

In this chapter, we discuss the techniques available to execute a quantitative analysis, factors that validate the decision to use bibliometrics, some works presented in this research field, their contributions, and our contributions.

Literature reviews are established to help academics deal with remarkable scientific knowledge production volumes (ÖZTüRK; KOCAMAN; KANBACH, 2024). It provides us ways to examine the status of a research field, categorize knowledge, and help researchers observe the efforts presented in certain research topics while giving more insight on characteristics presented in determined studies (ÖZTüRK; KOCAMAN; KANBACH, 2024). With large quantities of studies, the necessity to audit research from time to time becomes more relevant since we want to select works for examination and identify strengths and flaws in research topics (ÖZTüRK; KOCAMAN; KANBACH, 2024). These frameworks became popular with businesses by exploring data analytics to maximize the production process (ALSOLBI et al. 2022). NPOs also use it to monitor, evaluate, and determine barriers to their success and can provide meaningful visualizations to support decision-makers (ALSOLBI et al. 2022). The fact is that data analytics has become a pivot for everyone, and for that, we have frameworks that focus on different aspects of literature production, each one with limitations that will be synthesized in this section.

Systematic reviews aim to provide a comprehensive, unbiased synthesis of many relevant studies in a single document (AROMATARIS; PEARSON, 2014). This type of analysis tends to be focused on a specific topic, often using manual procedures to better comprehend the field in question and produce well-condensed results (DONTHU et al. 2021a). Formal approaches to systematic review in the field have often focused on applied questions (GRAMES et al. 2019).

Meta-analysis is a reliable method for exploring empirical evidence of relationships between variables while uncovering relationships not studied in existing research (AGUINIS et al. 2011) (DONTHU et al. 2021a). This technique is recommended when the review focuses on summarising results rather than engaging deeply with the content. It requires a sufficient quantity of homogeneous studies to justify and sustain this approach (DONTHU et al. 2021a).

Bibliometrics analysis is a powerful tool that allows us to summarise large amounts of data, helping us understand the nuances in the target field (DONTHU et al.

2021a). In a world where large quantities of documents are available online, this approach provides an advantage to those who want to understand certain aspects presented in large data sets. From business to information systems, bibliometric techniques have become popular given these circumstances, reflecting their suitability for handling large volumes of scientific data and producing high research impact (DONTHU et al. 2021a). This methodology is broken into two approaches: performance analysis and science mapping. Performance analysis is a quantitative method often used to evaluate journal productions by calculating metrics related to authors and publications. At the same time, science mapping explores a qualitative aspect by relationships established by intellectual and structural connections among research constituents (DONTHU et al. 2021a).

In the figure 2.1, we provide a table with a comparison of major review methods, contrasting them with the main goal of each one and the adequate conditions to use them more effectively:

Table 2.1: Comparison between methods of literature review

Comparison of major review methods.

Review type	Goal	When to use	When not to use	Scope	Dataset	Analysis
Bibliometric analysis	• Summarizes large quantities of bibliometric data to present the state of the intellectual structure and emerging trends of a research topic or field.	• When the scope of review is broad. • When the dataset is too large for manual review.	• When the scope of review is specific. • When the dataset is small and manageable enough that its content can be manually reviewed.	• Broad	• Large	• Quantitative (evaluation and interpretation) • Qualitative (interpretation only)
Meta-analysis	• Summarizes the empirical evidence of relationship between variables while uncovering relationships not studied in existing studies.	• When the focus of review is to summarize results rather than to engage with content, which may be broad or specific. • When studies in the field are homogenous. • When the number of homogeneous studies available is sufficiently high. • When the number of homogeneous studies remaining after removing low quality studies is sufficiently high.	• When studies in the field are heterogeneous. • When the number of homogenous studies is relatively low. • When the number of high-quality homogeneous studies is relatively low.	• Broad • Specific	• Large • Small but adequate	• Quantitative (evaluation and interpretation)
Systematic literature review	• Summarizes and synthesizes the findings of existing literature on a research topic or field.	• When the scope of review is specific. • When the dataset is small and manageable enough that its content can be manually reviewed.	• When the scope of review is broad. • When the dataset is too large for manual review.	• Specific	• Small	• Qualitative (evaluation and interpretation)

Source: Donthu et al. (2021a)

Donthu et al. (2021a) explain that meta-analysis helps us have an overall notion of the efforts, relationships, across-study variance, and the factors that explain the characteristics found in the group of studies selected. At the same time, systematic reviews such as domain-, method-, and theory-based reviews encapsulate the acquisition, arrangement, and assessment of the extant literature, usually by making a manual procedure, giving us a more restricted scope. While meta-analysis and bibliometric analysis are quite similar when we look at the quantitative aspect, the focus of meta-analysis is to summarise empirical evidence while looking for relationships between variables, usually helping to clarify mixed empirical evidence while extending the topic in observation (DONTHU et al. 2021a). On the other hand, bibliometric analysis is used to summarise a field's bibliometric and intellectual structure by analyzing the social and structural relationships between different research constituents (e.g., authors, countries, institutions, topics) (DONTHU et al. 2021a).

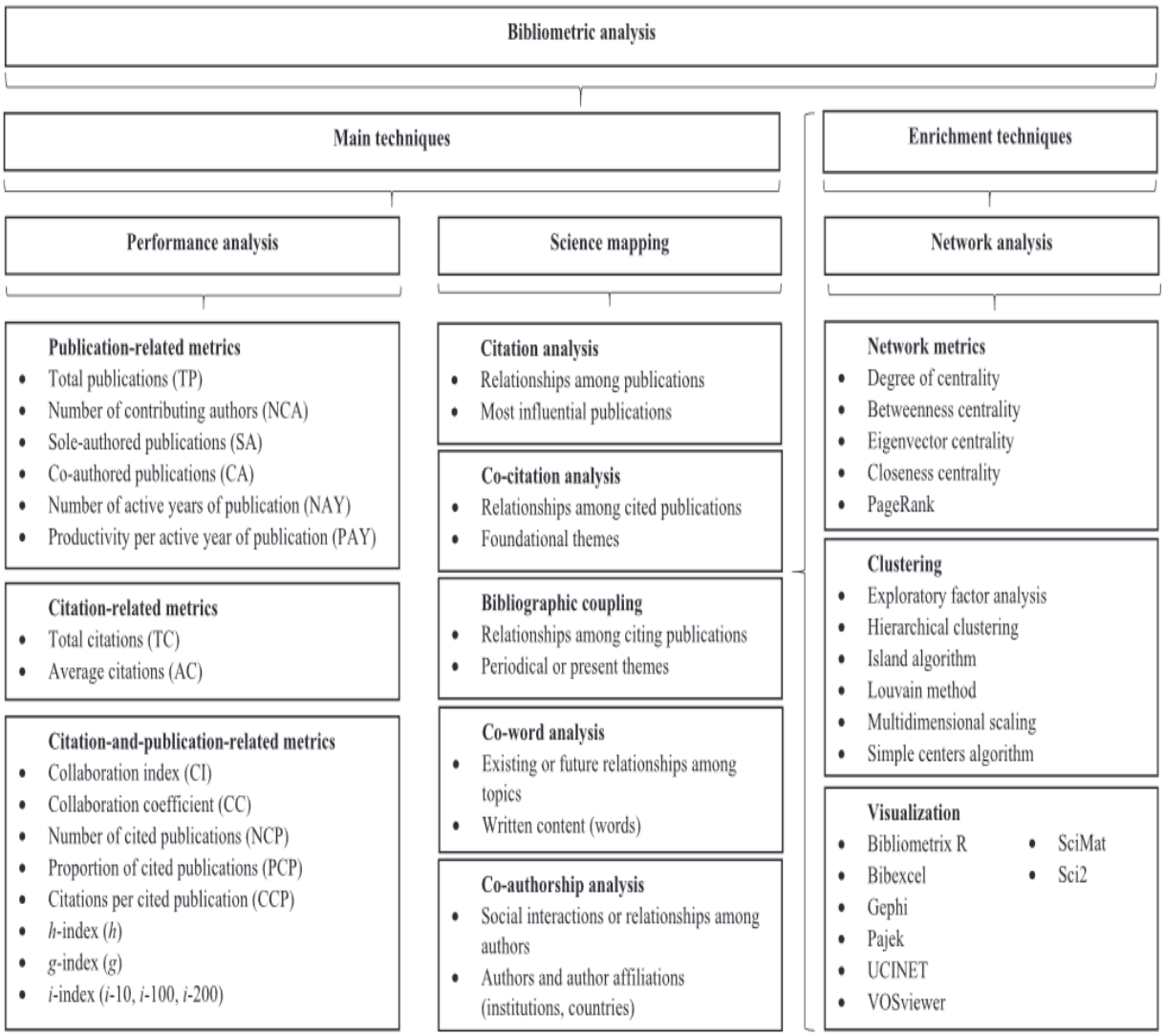
For our purposes of evaluating the growth aspect of RITA while dealing with a large dataset of scientific studies and observing the relationships developed over those 16 years, bibliometric analysis is the best fit to help us in this investigation. Research of this type has been used to evaluate different types of documents, including newspapers, social media, and scientific journals (DONTHU et al. 2021a). For example, Verma and Gustafsson (2020) focused on the booming trend of COVID-19 studies to observe trending topics in scientific production. In Donthu et al. (2021b), the goal was to evaluate the production aspect of journals by calculating bibliometric indices, mapping the geographical location of authors, and observing popular keywords. Ellegaard and Wallin (2015) extracted a dataset from the Web Of Science and compared the impact of Library and Information Sciences (LIS) articles to non-LIS ones (applied and subject-based studies).

In all of the studies presented above, despite minor differences in the target or approach, those focused on evaluating production value tend to concentrate only on published articles since it's the core aspect when assessing a journal's productivity. In comparison, the present work aims to provide a comprehensive overview of RITA as a scientific journal by examining the productivity aspect and analyzing rejected articles to contrast with accepted ones. This approach helps us understand key differences and the volume of authors and institutions rejected by the journal during its existence.

Our main goal is to evaluate the RITA profile as a journal using bibliometrics to observe production value for both authors and publications perspectives while exploring

the subject pres must follow the guideline sections formed during his existence between authors and institutions by using techniques from both performance analysis and science mapping toolbox as well. The complete toolbox overview is shown in the figure 2.1:

Figure 2.1: The Bibliometric Toolbox



Source: Donthu et al. (2021a)

For this study, techniques related to publication and citation metrics are used on the performance side, and for science mapping, the choice is citation, co-word, and co-authorship analysis. These subsets are defined to evaluate the journal productivity while observing RITA network expansion across time to see how far RITA has reached since its conception. To achieve this, we need to follow the guidelines defined to execute this procedure correctly. The basic steps presented by Donthu et al. (2021a) to make a

bibliometric analysis are:

- Step 1: Define the aims and scope of the bibliometric study.
- Step 2: Choose the techniques for bibliometric analysis.
- Step 3: Collect the data for bibliometric analysis.
- Step 4: Run the bibliometric analysis and report the findings.

The goals for steps (1) and (2) are already defined above; for (3) and (4), an extraction of datasets from the RITA submission register was made and combined with additional information to ensure that we choose in step (2) to execute (4) as intended. The chapter 3 presents in-depth how to fulfill these steps using a CRISP-DM approach.

3 METHODOLOGY - CRISP-DM FRAMEWORK

CRISP-DM is an industry-independent process model for data mining. This method has six iterative phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (SCHRöER; KRUSE; GóMEZ, 2021). This framework provides a robust life cycle process that can be reused as needed, making it a powerful tool for evaluating RITA production each year. Each phase is going to be discussed and contextualized in the next section.

3.1 Problem Understanding

The business situation should be assessed to get an overview of the available and required resources. Determining the main goal is one of the most important aspects of this phase. First, the data mining type should be explained (e. g. classification) and the data mining success criteria (like precision). A compulsory project plan should be created (SCHRöER; KRUSE; GóMEZ, 2021).

In this case, RITA is a scientific open journal that publishes one volume with at least two numbers per year. It wants to evaluate its production to understand its quality and impact and plan new ways to improve the general process. To help with this, the main goal is to define a model that supports qualitative and quantitative analysis to ensure the editorial team has a broad overview of his production. To achieve this goal, the extraction of bibliometric indexers with a text-mining approach is combined to explore more aspects found in the journal dataset.

3.2 Data Understanding

Collecting data from data sources, exploring and describing it, and checking the data quality are essential tasks in this phase. To make it more concrete, the user guide describes the data description task using statistical analysis and determining attributes and their collations (SCHRöER; KRUSE; GóMEZ, 2021).

Since we want to evaluate data from published studies and rejected articles, we need to gather information stored in the submission process to have these two categories available. The SEER system, hosted by the university and used by RITA, has a database

tracking the submission process. The administration panel can access this data used to manage the journal.

Two datasets were given by the editor-in-chief for this study, containing entries of all articles submitted from 2008 to 2023, with 1174 articles. The first dataset includes articles with an identifier number and classifies them under one of the following statuses: published, declined, review, copy-editing, or submission. Each entry in this dataset contains information about the title and abstract, the names of authors and co-authors, Open Researcher and Contributor IDs (ORCIDs), emails, biographical statements, dates of submission and publishing, the Digital Object Identifier (DOI) of each article, and the current status of the publication. The second dataset includes a record of the number of views registered per article.

3.3 Data Preparation

Data selection should be conducted by defining inclusion and exclusion criteria. Bad data quality can be handled by cleaning data. Based on the model used (defined in the first phase), derived attributes must be constructed. For all these steps, different methods are possible and are model dependent (SCHRöER; KRUSE; GóMEZ, 2021).

Our approach consisted of extracting these datasets from the journal's administrator panel and establishing a cleaning process to remove incomplete or irrelevant data for the analysis. First, we removed all columns that were not going to be used for the analysis (e.g., emails, biographical statements, ORCIDs), and then we deleted all incomplete rows. The second part involved removing special characters in the abstracts that could have interfered with data processing.

RITA receives articles in three different languages: Portuguese, English, and Spanish (<https://seer.ufrgs.br/index.php/rita/about>). Therefore, it was necessary to normalize the language between articles to concentrate the text-mining results. For that, we upload the dataset provided in a Comma Separated Values (CSV) format to Google Sheets and apply a formula that translates the abstracts from any language to English (e.g. `=IFERROR(GOOGLETRANSLATE(C2; "auto"; "en"))`), enabling us to use them for Natural Language Processing (NLP) text classification.

At this point, we also split the data into two categories: accepted articles and rejected ones. This decision was made to simplify the subsequent steps involving web scraping to gather additional information available for published studies.

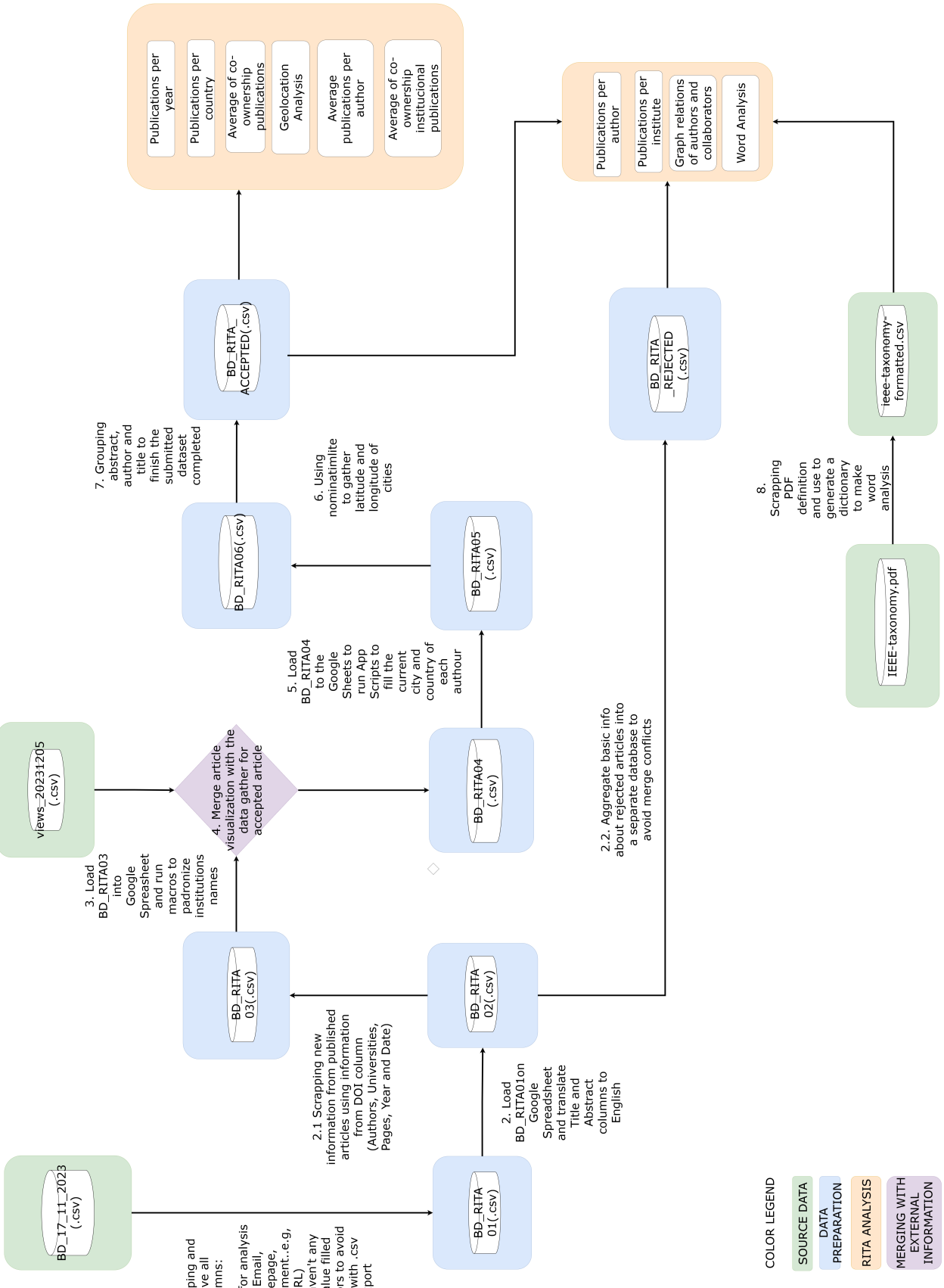
Afterwards, we merged the views dataset with the dataset of published articles. Additionally, the Google Apps Script platform was used to develop scripts in JavaScript language to fetch through the dataset, identify the name of the institution related to each author, and fill in the information related to the city and country related to them.

3.4 Modelling

The data modelling phase involves selecting the adequate technique and building the test case and the model. All data mining techniques can be used. The choice generally depends on the business problem and the data. Specific parameters have to be set to build the model. For assessing the model it is appropriate to evaluate the model against evaluation criteria and select the best ones (SCHRöER; KRUSE; GóMEZ, 2021).

To achieve this goal, we implemented a pipeline derived from a previous exploratory analysis executed by Villalobos-Cid (2022), maintaining the same purpose of refining data for analysis but with additional phases to include rejected data categorization and processing in the schema. The figure 3.1 provides us with an overview look:

Figure 3.1: Pipeline architecture to process RITA dataset



Source: The Author

This pipeline was made using R language, and the description of each step enumerated in figure 3.1 is provided down below:

1. The first part of the process was the data cleaning. We removed all of the useless information and rows with incomplete information (See Figure 3.1, item 1).

2. In the second step, we made a manual upload of the dataset to Google Sheets. We ran a spreadsheet formula to translate all the abstract columns to English and split the database into two subsets to avoid merging problems with web-scraped information and the view count table. Here is what we did with each database (See Figure 3.1, item 2)..

    2.1. Grouping data from published articles and using a web scrapping process to gather information crucial to bibliometrics, such as co-authorship and the number of pages of each article (See Figure 3.1, item 2.1)..

    2.2. We grouped data from published articles and used a web scraping process to gather information crucial to bibliometrics, such as co-authorship and the number of pages of each (See Figure 3.1, item 2.2).

3. After generating the new intermediate dataset, we uploaded it again to Google Sheets to use scripts created in Google Apps Script to define a pattern for institution names and concentrate the total number of institutions without losing information (See Figure 3.1, item 3).

4. We decided to download documents and merge the view count information into our principal dataset using R language so we could use this information to explore different aspects explained in the subsequent topics (See Figure 3.1, item 4)..

5. After the merge, we uploaded the result to Google Sheets to run another Google Apps Script to fill in information related to the city and country based on the institution's name (See Figure 3.1, item 5)..

6. With the city location of each author identified, we used an R language function contained in the `nominatlimit` package called `geo_lite` to process each city name and store the latitude and longitude of the city (See Figure 3.1, item 6).

7. Here, we combined the missing data stored in the subsequent datasets to group all useful information into one dataset called BD_RITA_ACCEPTED (See Figure 3.1, item 7).

8. In this last step, we process scrapped data from IEEE_Thessarus to generate a valid dictionary for text processing (See Figure 3.1, item 8).

As we previously mentioned, we used web scrapping techniques in steps 2 and 8 (Figure 3.1) to gather vital information to execute bibliometric analysis and have a keyword-analysis dictionary. We used the web scrapper library `rvest` to make HyperText Markup Language (HTML) parsing from RITA website, and `pdftools` library to split the IEEE_Thessarus definitions into a Comma-Separated Values (.csv) and with them mount a script to generate a dictionary with areas and subareas, allowing us to filter studies by using a formal dictionary with definitions made by IEEE.

3.4.1 Bibliometrics Gathered

We've used the final dataset to compute all indexes available related to publications and authors present from 2008 to 2023. Now, we going to explain each of the bibliometric indexers calculated in this research:

3.4.1.1 Publication Indexers Related to Articles Metrics

- Annual Growth Rate of Publications (AGR)

  Annual growth rate (AGR) is the change in the value of a measurement over a year (GUPTA; HASAN, 2018). To compute this, we used the formula down below:

$$AGR = (\frac{\#\text{Manuscripts}[a] - \#\text{Manuscripts}[a-1]}{\#\text{Manuscripts}[a]} \times 100)$$

  Where a represents the year desired for the calculus.

- Cumulative Annual Growth Rate (CAGR)

  It corresponds to the cumulative growth rate of the number of publications over a specific period (KULKANJANAPIBAN; SILWATTANANUSARN, 2021). To compute this, we used the formula down below:

$$\text{CAGR} = \left(\frac{\#\text{Manuscripts}_{\text{CUM}}[a]^{\left(\frac{1}{\text{year}(a)-1}\right)}}{\#\text{Manuscripts}[a]} \right) \times 100$$

  Where a represents the year desired for the calculus.

- Relative Growth Rate (RGR)

  It corresponds to the increase in the number of articles published per unit of time (KUMAR; KALIYAPERUMAL, 2015) (RATHIKA; THANUSKODI, 2021). The average RGR for a time interval can be calculated as follows below:

$$\text{RGR} = \frac{\ln(\#\text{Manuscripts}[a]) - \ln(\#\text{Manuscripts}[a-1])}{\text{year}(a) - \text{year}(a-1)}$$

  Where a represents the year desired for the calculus.

- Doubling Time (DT)

  There is a direct relationship between RGR and the time it would take for the number of publications to double. To calculate the doubling time (DT), a standard $\ln(2)$ represents the double publication in time equation (approximate 0.693) (KUMAR; KALIYAPERUMAL, 2015) is used here (RATHIKA; THANUSKODI, 2021), resulting in the following equation:

$$\text{DT} = \frac{0.693}{\text{RGR}}$$

3.4.1.2 Publication Indexers Related to Authors Productivity

- Author Productivity Per Publication (AAPP) and Productivity Per Author Associated (PPAA)

  These indicators included in (Gupta and Hasan 2018) measure the relationship between the number of authors and publications.

$$\text{AAPP} = \frac{\#\text{Authors}[a]}{\#\text{Manuscripts}[a]}$$

$$\text{PPAA} = \frac{\#\text{Manuscripts}[a]}{\#\text{Authors}[a]}$$

  Where a represents the year desired for the calculus.

- Degree of Collaboration (DC)

  It is defined as the relationship between articles with more than one author (Nm) and those written by only one person (Ns) (SAVANUR; SRIKANTH, 2010) (GUPTA;

HASAN, 2018) (BARIK; JENA, 2019) (DAS; KAUR; DR, 2021).

$$\text{DC} = \frac{N_m}{N_m + N_s}$$

- Collaborative Index (CI)

  It is a weighted average of authors per joint article. That is, those who have more than one author (SAVANUR; SRIKANTH, 2010) (GUPTA; HASAN, 2018) (BARIK; JENA, 2019) (DAS; KAUR; DR, 2021).

$$\text{CI} = \frac{\sum_{j=1}^{A} j * f_j}{N}$$

  Where

  - j is the number of authors per article, for example, 1, 2, 3, ... up to A.
  - fj is the number of articles with j authors.
  - N is the total number of articles published that year.

- Collaborative Coefficient (CC)

  Collaborative coefficient is a measure of collaboration in research, that reflects both the mean number of authors per paper as well as the proportion of multi-authored papers. Although it lies between the values 0 and 1, and is 0 for a collection of purely single-authored papers, it is not 1 for the case where all papers are maximally authored, i.e., every publication in the collection has all authors in the collection as co-authors (SAVANUR; SRIKANTH, 2010).

$$\text{CC} = 1 - \frac{\sum_{j=1}^{A} \frac{1}{j} * f_j}{N}$$

  Where

  - j is the number of authors per article, for example, 1, 2, 3, ... up to A.
  - fj is the number of articles with j authors.
  - N is the total number of articles published that year.

- Modified Collaboration Coefficient (MCC)

  This indicator is a variation of the previous indexer explained above. MCC considers the frequency of collaborations and the total number of authors involved

in a particular article or scientific work. (DAS; KAUR; DR, 2021).

$$\text{MCC} = (\frac{N}{N-1})(\frac{\sum_{j=1}^{A} \frac{1}{j} * f_j}{N})$$

Where

- j is the number of authors per article, for example, 1, 2, 3, ... up to A.
- fj is the number of articles with j authors.
- N is the total number of articles published that year.

- Average number of pages per article (PNA)

  Corresponds to the average number of pages per published article.

$$\text{PNA} = \frac{\sum \text{\#Number\_pages}}{\text{\#Manuscript\_by\_year}}$$

3.4.1.3 Indexers related to the view count

- Average of views per year (VPA)

  The average of views given a certain year can be computed with the equation below:

$$\text{VPA} = \frac{\#\text{Views}[a]}{\#\text{Manuscripts}[a]}$$

  Where a is the year desired for the calculus

- Total visits adjusted per articles (VAPA)

  This version is to compute the average adjusting a year gap to normalize the value:

$$\text{VAPA} = \frac{\frac{\#\text{Views}[a]}{\#\text{Manuscripts}[a]}}{\text{Final}_{\text{year}} - \text{Current}_{\text{year}} + 1}$$

3.4.2 Evaluation

The results are checked against the defined business objectives in the evaluation phase. Therefore, the results must be interpreted, and further actions must be defined. Another point is that the process should generally be reviewed (SCHRöER; KRUSE; GóMEZ, 2021).

3.4.2.1 Text Processing

The objective of obtaining information from the articles to have a look at the subjects researched by the authors is realized by text mining the abstract registered in the database, filtering them to obtain only the possible candidates for keywords, and using the dictionary mounted using IEEE Thesaurus definitions to standardized the topics founded. After categorizing and quantifying the topics and subtopics found in abstracts, a function called `textstat_keyness` from the `quanteda` package that applies a χ^2 test that contrasted two sets of words to visualize the relevance of the target dataset against a reference provided.

3.4.2.2 Visualization Analysis

For a more comprehensive view, we are going to have three types of graphs to give the reader a better look at certain characteristics

- Vertical bar charts for looking at the volume of RITA production.
- Graphs when authors are the nodes and the vertices are the connections established between them. In this category, we have graphs to observe both the author's connectivity and geographical localization by pinpointing the authors using latitude and longitude collected before.
- Horizontal bar charts to observe the impact of keyword analysis in contrast to a given reference.

3.4.3 Deployment

The proposed solution above was developed in R programming language and divided into specialized scripts for each objective. The packages list used is: for ordering data was `dplyr`; for graphics is `ggplot2`, `quanteda`, `tidyverse`, and `leaflet`; for web scrapping, we used both `rvest` and `pdftools`; for gather latitude and longitude was `nominatimlite`; we also used `stringr` for manipulation of the abstracts; and for the tables we used `knitr`.

First, we have to extract the datasets accessed only by the administration panel of RITA (<https://seer.ufrgs.br/rita>). We executed a cleanup script to remove undesired/useless information. After that, we parsed this dataset with content gathered

from RITA digital editions by fetching information on each article through the volumes using the DOI number registered in the dataset. After uploading our new intermediate dataset into Google Sheets, we apply the formula to translate every abstract to English so we can have a normalized version for later usage. To ensure that the next steps will not cause any major issues, we split the dataset between articles accepted and rejected to avoid merging conflicts with the views dataset and merge the accepted ones with that information, making the joins based on the identification (doc) column. For the last steps, scripts were created with Google App Script, a resource from Google that allows writing macros for spreadsheets using JavaScript to read the name of every author institution and search for the city and country location. To finish the dataset of accepted articles, a manual merge of the columns missed during the view count merge was made using Google Spreadsheets.

The indexers to evaluate the journal are calculated per year, using the formulas defined in the item 3.4.1.1. To ensure that our results are valid, we perform statistical tests on our samples to ensure that our dataset is significant enough. These results are grouped per year and exhibited in tables. We also want to look at authors and institution presence in some categories, such as publications, connections, and article visualization. For that, we count the frequency registered by each category and show them in tables to measure these while found inside by looking for patterns in these results. For the graphical part, we generate results to observe the volume of submissions received, draw geo-located graphs to evidence the location of submissions, and create relationship graphs based on co-authorship to evidence the network constructed by authors and institutions with the passage of the years. We generated horizontal graphs from word analysis by extracting keywords from the abstracts, classifying them using the dictionary created with the IEEE Thesaurus definition, and applying frequency usage tests by looking at frequent terms contrasting with other years and comparing them with rejected articles.

All the scripts used to establish the pipeline along with the analysis executed are available on Github <https://github.com/cgsrjunior/bibliometric-rita>, along with a basic description of each script on the README.md but the datasets cannot be provided along with the code since the data is not in public domain.

4 EXPERIMENTAL RESULTS

In this chapter, we're going to explore our results while addressing the Research Questions proposed at section 1.5

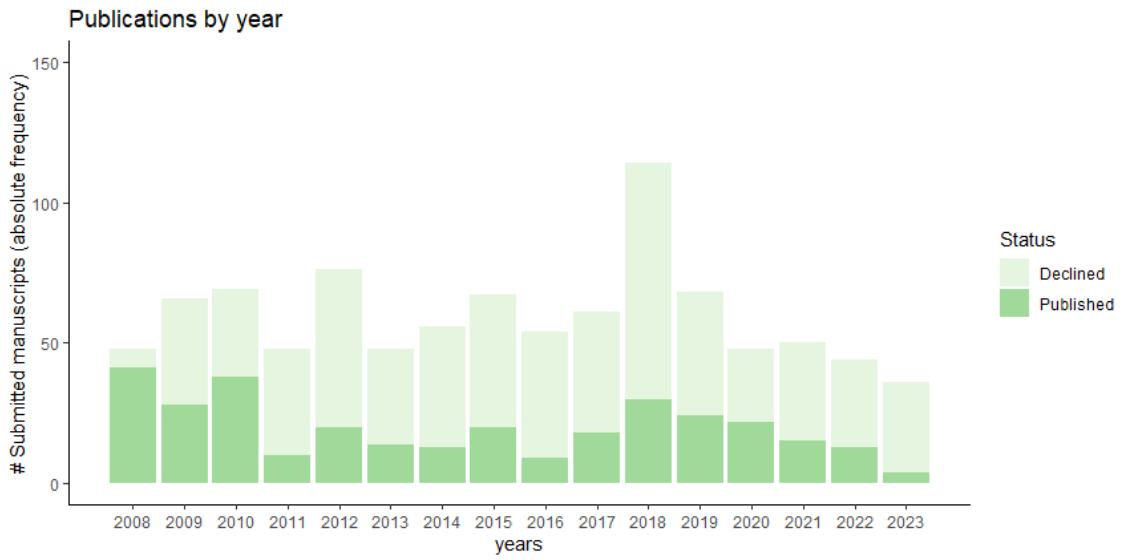
4.1 How has the number of publications and the publication trends evolved over the past 15 years, and which years had the highest and lowest number of publications?

In this phase, we want to concentrate all the available information and aggregate it by year to observe the brute volume of submissions registered in the dataset. The sections below show this.

4.1.1 Publication Trends and Evolution

First of all, we want to look at all submissions received and processed by RITA, from 2008 to 2023 to see if the sheer volume of submissions and publications has grown over time:

Figure 4.1: Graph with all articles Published and Declined in RITA database from 2008 to 2023



Source: The Author

Table 4.1: Table With Articles Published and Declined in RITA database from 2008 to 2023

Year	*Total articles*	*Published*	*Declined*
2008	48	41	7
2009	66	28	38
2010	69	38	31
2011	48	10	38
2012	76	20	56
2013	48	14	34
2014	56	13	43
2015	67	20	47
2016	54	9	45
2017	61	18	43
2018	114	30	84
2019	68	24	44
2020	48	22	26
2021	50	15	35
2022	44	13	31
2023	36	4	32
Total	**973**	**319**	**674**

Source: The Author

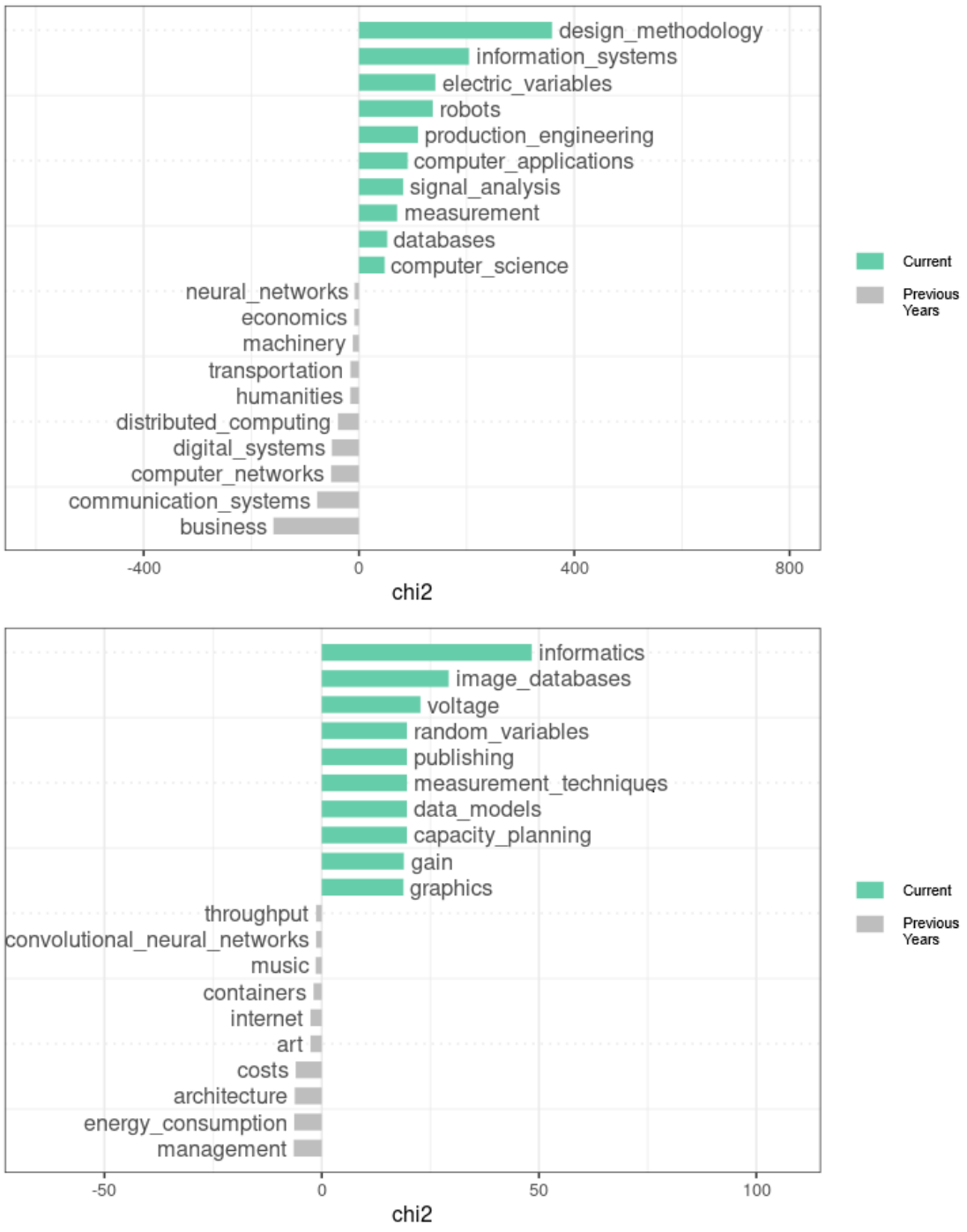
In the figure 4.1 we can observe that the number of submissions have small variations between 2008 and 2017, with a great volume of submissions in 2018. But after that, we have a decline in the registered submissions that extends from 2020 to 2023. The year with the highest number of articles received is 2018, and the last one is 2023.

By looking at the number in table 4.1, we can notice that only approximately 32.79% of articles received by RITA are published, a factor that evidence that despite the increase in received submissions. Since RITA is focused only on computer science topics and has strict selection criteria, this means that we have received a great amount of low-quality studies. We can pinpoint that fact due to the Qualis of RITA (B3), which can indicate to newcomers in research that it's a good place to try their first publication.

For the last, we want to observe all topics and subtopics presented during this period to see the evolution of computer science-researched subjects in RITA. The graphs below show the evolution of keywords of areas and subareas from 2008, 2013, 2018, and

2023 to have a snapshot from different periods and see the change of focus in each period. Our test consists of analyzing the distribution of keywords identified in the abstracts by using a Thesaurus dictionary and seeing the distinct of the most used terms in the selected years in contrast with terms used previously. The range of χ^2 in the graphs indicates the frequency of terms found in each group. The group on the right represents the keywords found in 2008 and on the left, we have keywords from previous years.

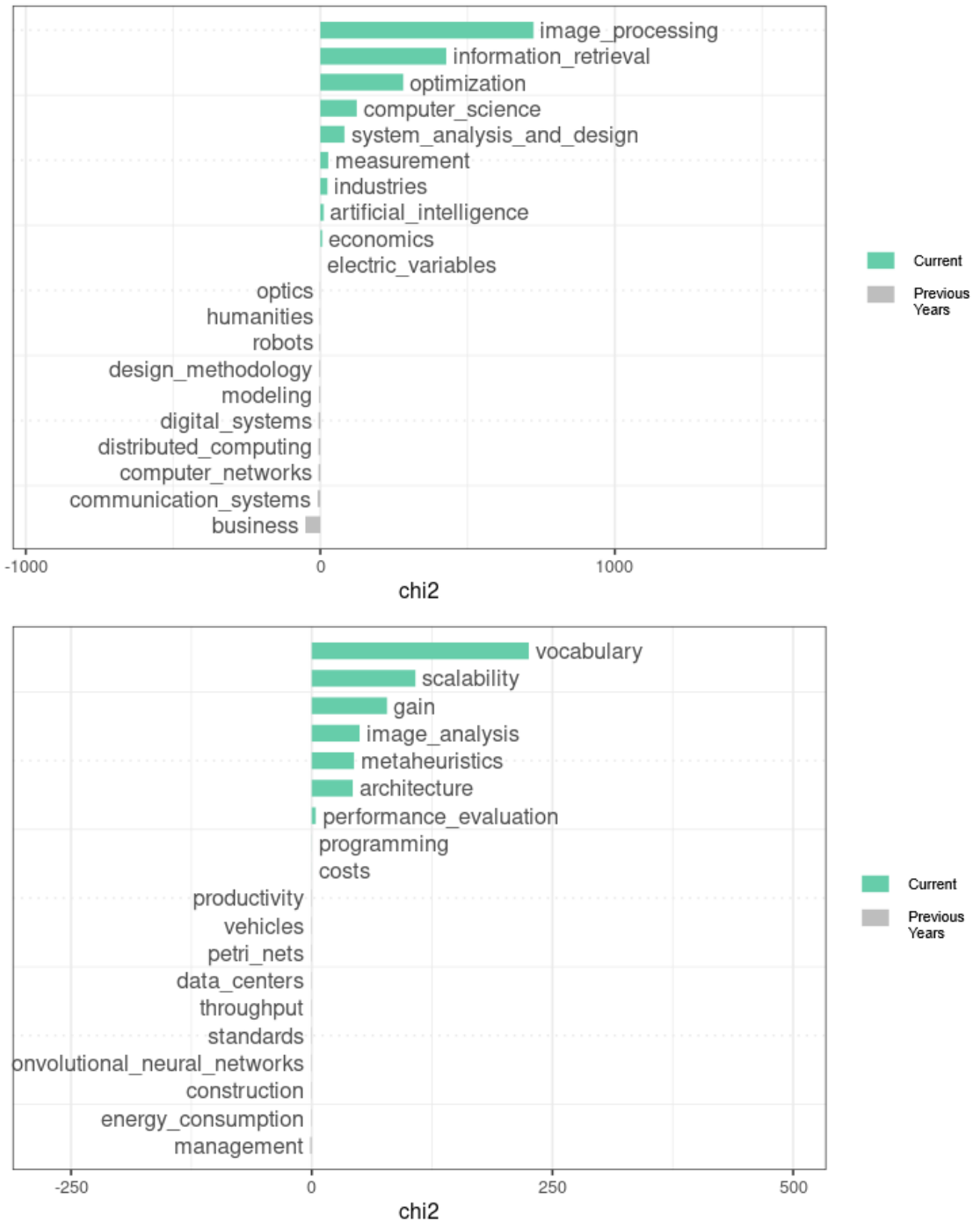
Figure 4.2: Distinction of Areas and Subareas Keywords in 2008 and Contrast With Previous Years



Source: The Author

In the figure 4.2 we can observe the tendencies of research can be mainly attributed to data analysis and computer engineering topics.

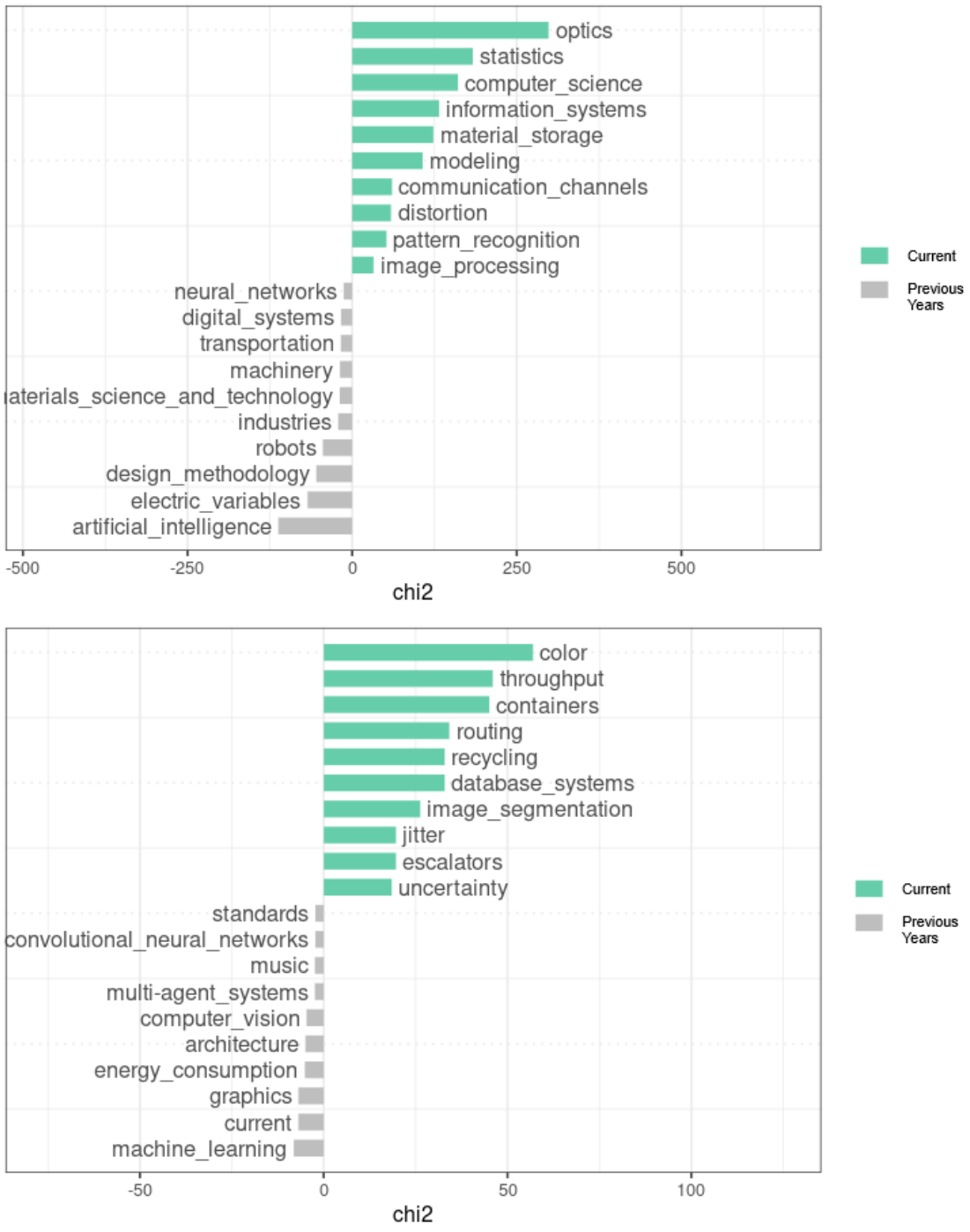
Figure 4.3: Distinction of Areas and Subareas Keywords in 2013 and Contrast With Previous Years



Source: The Author

In figure 4.3 the focus appeared to shift and became more distributed than in 2008, which can be grouped the results into image manipulation, optimization problems, and hardware performance.

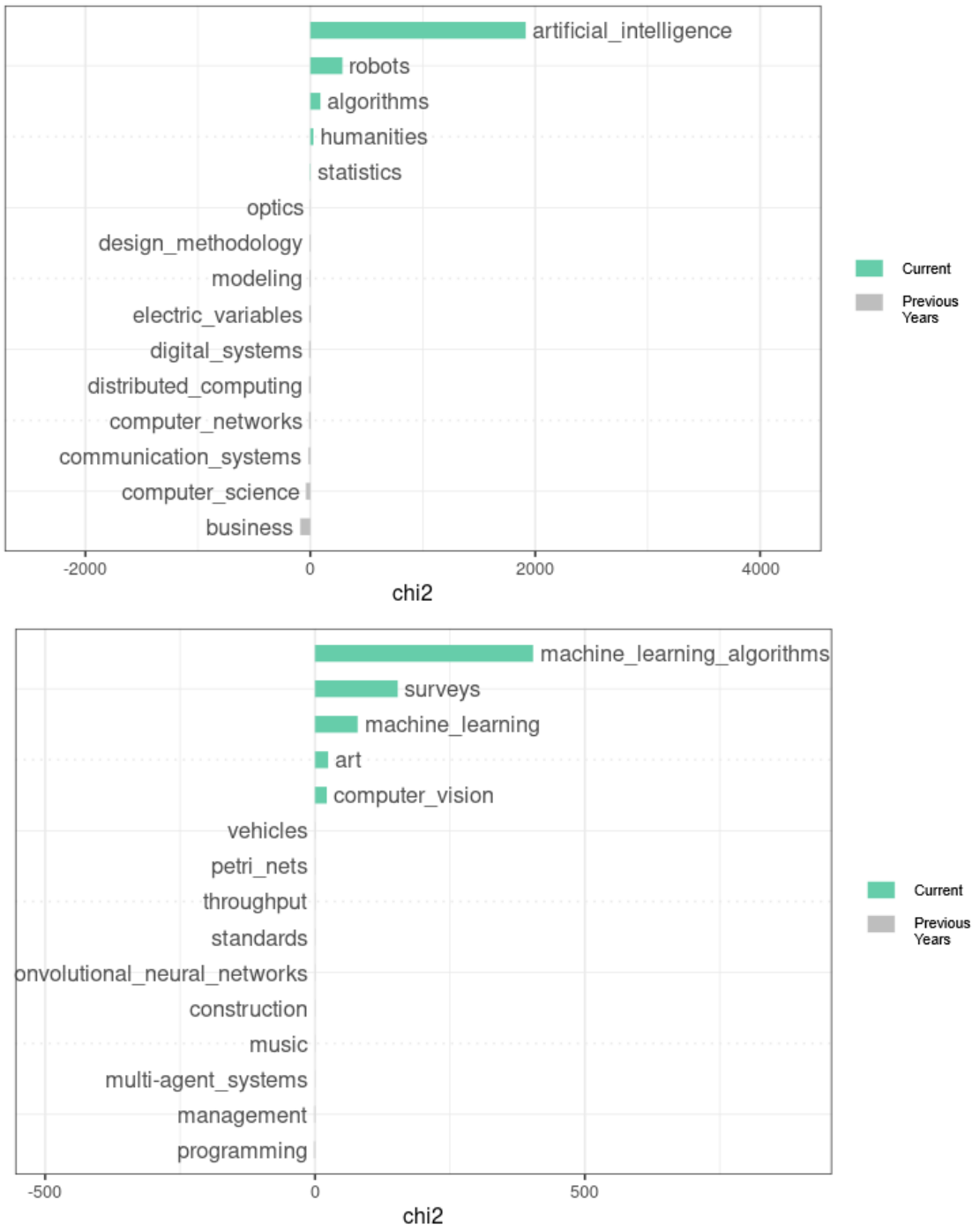
Figure 4.4: Distinction of Areas and Subareas Keywords in 2018 and Contrast With Previous Years



Source: The Author

In figure 4.4 we still have image manipulation present but on a minor scale if we compare it to 2013. The new topics that dominated 2018 can be classified as communication engineering and information systems.

Figure 4.5: Distinction of Areas and Subareas Keywords in 2023 and Contrast With Previous Years



Source: The Author

In the figure 4.5 we can notice the current trend of AI dominating all categories, with machine learning leading them but we also can see the beginning of generative AI image studies appearing on RITA.

The terms identified in each year mostly follow the trends emerging in computer

science. In 2008, we have a shift in smartphones industry, with new technologies rising and the data science gaining popularity as a career among professionals. As for 2013, the proliferation of big data takes the main stage, exploring different aspects, such as information retrieval and image processing. Looking at 2018 is notable that the demand for the software engineering increased and with them the ascent of works that involves information systems and network communication. And for the last in 2023, the booming trend of AI can be observed as well, with the major of studies talking about machine learning, the hot topic that keeps growing today (SARKER et al. 2021) (JAMIE FOSTER SCIENCE, 2023).

4.2 How do different bibliometric indices assess the quality and impact of scientific publications in RITA over the last 15 years?

In this section, we want to calculate the bibliometric indexers to see how our journal has progressed along its existence.

4.2.1 Bibliometric Indices Analysis

First, we start by computing the publications indexes for each year, from 2008 to 2023.

Table 4.2: Table bibliometrics indexes (publications) computed from 2008 to 2023

Years	*Freq*	*Freq_AC*	*AGR*	*CARG*	*RGR*	*DT*
2008	41	41	0.00	0.00	0.00	0.00
2009	28	69	-31.71	146.43	0.52	1.16
2010	22	91	-21.43	103.38	0.28	2.15
2011	10	101	-54.55	116.16	0.10	6.03
2012	20	121	100.00	56.83	0.18	3.35
2013	14	135	-30.00	57.34	0.11	5.48
2014	13	148	-7.14	49.99	0.09	6.70
2015	20	168	53.85	35.53	0.13	4.64
2016	9	177	-55.00	45.12	0.05	12.06
2017	18	195	100.00	30.31	0.10	6.03
2018	29	224	61.11	22.68	0.14	4.31
2019	24	248	-17.24	23.65	0.10	6.03
2020	17	265	-29.17	25.72	0.07	8.61
2021	13	278	-23.53	26.57	0.05	12.06
2022	13	291	0.00	24.86	0.05	12.06
2023	4	295	-69.23	33.20	0.01	60.30

Source: The Author

These results show a decline in the average Annual Growth Rate (AGR) of -33.90% and an increase in Doubling Time (DT) on an average of 9.20 years, confirming suspicion of the journal's publication decline rate. These results also can represent a stagnation on the h-index[1] and g-index[2] of the journal since this generate an impact on the number of articles available, with decreases the probability of citation.

After that, this part of the section focuses on computing indexes related to the author's productivity, and what we found was registered in table 4.3:

[1]H-index measures the production and impact of a researcher or group of researchers (all being evaluated equally, whatever their rank on the publication)(HIRSCH, 2005)

[2]G-index is the (unique) largest number such that the top g articles received (together) at least g^2 citations(MINGERS; LEYDESDORFF, 2015)

Table 4.3: Table bibliometrics indexes (articles) computed from 2008 to 2023

Years	*Freq*	*AAPP*	*PPAA*	*DC*	*CI*	*CC*	*MCC*	*PNA*
2008	94	2.10	0.48	0.76	2.29	0.45	0.46	19.35
2009	101	3.46	0.29	1.00	3.61	0.68	0.71	20.53
2010	64	2.86	0.35	0.77	2.91	0.52	0.54	14.02
2011	30	3.00	0.33	0.80	3.00	0.54	0.60	19.33
2012	60	2.85	0.35	0.95	3.00	0.62	0.66	23.82
2013	33	2.29	0.44	0.79	2.36	0.48	0.52	21.30
2014	40	3.08	0.32	0.92	3.08	0.61	0.66	20.75
2015	60	2.95	0.34	0.85	3.00	0.57	0.60	25.12
2016	32	3.56	0.28	1.00	3.56	0.67	0.75	22.66
2017	50	2.78	0.36	0.83	2.78	0.54	0.57	14.70
2018	90	3.00	0.33	0.90	3.10	0.59	0.61	13.28
2019	75	3.08	0.32	0.75	3.12	0.52	0.54	11.87
2020	45	2.65	0.38	0.88	2.65	0.54	0.58	13.91
2021	39	2.92	0.34	0.85	3.00	0.56	0.61	14.67
2022	41	3.08	0.32	0.85	3.15	0.59	0.64	10.95
2023	18	4.25	0.24	1.00	4.50	0.77	1.02	12.22

Source: The Author

Here, we notice that the average Author Productivity Per Publication (AAPP) is 2.41; the Production Per Author Associated (PPAA) average is 32%, suggesting a more distributed pattern of publication by the author viewpoint; the Degree of Collaboration (DC) index average is 5%, evidence of the collaboration between authors here is more decentralised. This became more apparent when we looked at the average Collaborative Index (CI) of 0.19%; Collaborative Coefficient (CC) of 4%; and Modified Collaboration Coefficient (MCC) of 4%. For the last, we want to look at metrics related to the view count of articles:

Table 4.4: Table related to view metrics from 2008 to 2023

Ano	*Count*	*VPA*	*VAPA*
2008	41	3428.27	214.27
2009	27	2602.93	173.53
2010	21	1306.48	93.32
2011	10	2866.10	220.47
2012	20	1254.90	104.58
2013	13	1315.85	119.62
2014	14	1668.86	166.89
2015	21	1834.71	203.86
2016	10	3669.80	458.72
2017	17	1272.88	181.84
2018	29	987.86	164.64
2019	24	700.62	140.12
2020	17	824.88	206.22
2021	13	268.38	89.46
2022	13	174.69	87.34
2023	4	33.00	33.00

Source: The Author

The table 4.4 shows us the arithmetic mean of Views Per Article (VPA) observed from 2008 to 2023 is 1555.02, and the mean of Views Adjusted Per Article (VAPA) is 171.89. Despite the down in the number of publications, RITA retains a good number of readers.

4.2.2 Sample Validation

To ensure that we computed these indexes with a significant sample, we performed a hypothesis on the sample to check if the mean used for the pages and the views was significant. Our null hypothesis (H0) is that the sample average wasn't different from the population sample. All tests in this section were two-tailed, using a confidence interval of 95% and a p-value of 1. The results of each average are in the tables 4.5, 4.6 and 4.7 down below.

Table 4.5: Table With One Sample T-Test For Publications

Average	*Standard Deviation*
54.5	24.76

Source: The Author

Table 4.6: Table With One Sample T-Test For Pages Per Publication

Average	*Standard Deviation*
17.49	8.14

Source: The Author

Table 4.7: Table With One Sample T-Test For Views Per Publication

Average	*Standard Deviation*
175.97	342.48

Source: The Author

The bibliometrics indexes of publications indicate that the journal growth isn't solid. In the last years, the relative growth was small, indicating a decrease in the search of RITA for publications. The journal didn't have a substantial rise in the last five years and this is extenuated by the Double Time results in the last three years (since we have only a fraction of studies from 2023, it's important to notice that we didn't have a complete panorama from 2023 due to the time of dataset extraction, in the first semester of 2023). From the author's perspective of bibliometrics, RITA shows that we have an expected number of authors per publication, with a strong collaboration but with a decrease in the average of pages, which could indicate a new tendency of short studies format. As for the view indexers, we can observe higher visualizations in certain years (2008, 2009, 2011, 2016), which indicates that in those years we have articles registered that captured more attention during the following years than the rest.

4.3 What are the most viewed topics and articles over the past 15 years, and what characteristics do these popular articles share?

On this question, we want to explore visualization information related to articles to check the popularity of topics registered each year.

4.3.1 Popular Topics and Articles

We examined snapshots from 2023, 2018, 2013, and 2008 to analyze the evolution of article search trends over time. This analysis involved extracting keywords from the abstracts of articles and categorizing them based on their view counts by summing views of articles to a topic if the keyword matches with the abstract, allowing us to observe how the prominence of certain terms in article descriptions has shifted across different periods.

Table 4.8: Table With View Count Per Topic On 2008

area	*views*
Artificial intelligence	430288
Electric variables	110438
Robots	101007
Design methodology	86295
Industries	72905
Computer science	44352
Materials science and technology	11124
Information systems	10980
Modelling	9779
Probability	9225
Total	**735693**

Source: The Author

Table 4.9: Table With View Count Per Topic On 2013

area	*views*
Computer science	40964
Electric variables	21870
Economics	19008
Industries	13080
System analysis and design	4232
Optimization	1842
Modeling	1738
Materials science and technology	872
Image processing	869
Total	**91475**

Source: The Author

Table 4.10: Table With View Count Per Topic On 2018

area	*views*
Economics	64704
Computer science	44143
Industries	24660
Business	23400
Optics	19674
Electric variables	16364
Modeling	6572
Pattern recognition	4182
Algorithms	3301
Total	**201252**

Source: The Author

Table 4.11: Table With View Count Per Topic On 2023

area	*views*
Robots	351
Artificial intelligence	104

Source: The Author

The trends observed in 2008 (table 4.8) are related more to automation in general; in 2013 (table 4.9), the predominant subject was related to modelling and optimising problems; in 2018 (table 4.10), we can notice more trends focused on the business side of computer science and in 2023 (table 4.11), the booming of AI papers takes the main stage.

Now, the idea is to observe the articles with more views and see if we note any characteristics related to them.

Table 4.12: Ten Articles With More Views

author	*title*	*institution*	*views*	*year*
ACPLFC	Uma Introdução às Support Vector Machines	INST001 - Brazil	51663	2008
MM	Tutorial: Introdução à Visão Computacional usando OpenCV	INST019 - Brazil	27564	2009
AB	QEDS: Um Simulador Clássico para Distinção de Elementos Quântico	INST004 - Brazil	25779	2016
LS	Estudo do Padrão Avançado de Criptografia AES – Advanced Encryption Standard	INST020 - Brazil	13048	2011
PBM	Teorias da Aleatoriedade	INST002 - Brazil	9708	2008
RP	Um Olhar Sociotécnico sobre a Engenharia de Software	INST021 - Brazil	9613	2008
JW	Processos de Decisão de Markov: um tutorial	INST006 - Brazil	9516	2008
SKG	A Gentle Introduction to Predictive Filters	INST006 - Brazil	9225	2008
RS	Java Advanced Imaging API: A Tutorial	INST030 - Brazil	8010	2008
HFM	Desenvolvimento e Avaliação de um Protocolo Eletrônico para Atendimento e Monitoramento do Paciente com Doença Celíaca	INST024 - Brazil	5946	2010

Source: The Author

Table 4.12 shows that our journal's most popular articles are those related to tutorials or introductions to certain topics, indicating that most people looking for articles to read in RITA are newcomers who want to learn about techniques to improve their abilities in computer science.

The view count per topic reveals that certain subjects from specific years receive more attention than others, likely due to the varying levels of interest and relevance of these studies over time. The most viewed articles in RITA suggest that our current audience may lack maturity, indicating a need to attract a more mature readership to encourage higher-quality submissions to the journal.

4.4 Which authors and institutions have the most published and rejected articles, and what is their collaboration network?

This question aims to provide a more in-depth look at the author's presence in RITA by exploring his production to get insights into the network created by the journal.

4.4.1 Authors, Institutions, and Collaboration Networks

In this subsection, we want to summarise the more frequent authors and institutions to see which entities are more present in the RITA lifespan.

Table 4.13: Institutions With More Entries On The Journal

Institutions	*Count*
INST001 - Brazil	75
INST002 - Brazil	59
INST003 - Brazil	46
INST004 - Brazil	34
INST005 - Brazil	27
INST006 - Brazil	26
INST007 - Brazil	25
INST008 - Brazil	24
INST009 - Brazil	24
INST010 - Brazil	22

Source: The Author

Table 4.14: Institutions With More Rejections On The Journal

Institutions	*Count*
INST009 - Brazil	16
INST011 - Brazil	12
INST012 - Brazil	12
INST001 - Brazil	9
INST013 - Brazil	6
INST014 - Brazil	5
INST015 - Brazil	5
INST016 - Brazil	5
INST007 - Brazil	5
INST018 - Brazil	4

Source: The Author

As we can observe in the table 4.13 great part of our production is from Brazil itself and more specifically, public universities. When we look at table 4.14, the most notable thing here is despite 2/3 of the submitted articles being rejected, the top 10 of rejected articles didn't concentrate many articles rejected, which indicates that we received articles from many different sources around the world.

Table 4.15: Authors With More Entries On The Journal

Author	*Institution*	*Count*
AFS	INST012 - Brazil	6
JPMO	INST002 - Brazil	6
AW	INST008 - Brazil	4
ETP	INST022 - Brazil	4
LAD	INST001 - Brazil	4
ASAN	INST023 - Brazil	3
GRAC	INST007 - Brazil	3
ITP	INST024 - Brazil	3
JSS	INST001 - Brazil	3
JK	INST003 - Brazil	3

Source: The Author

Table 4.16: Authors With More Rejection On The Journal

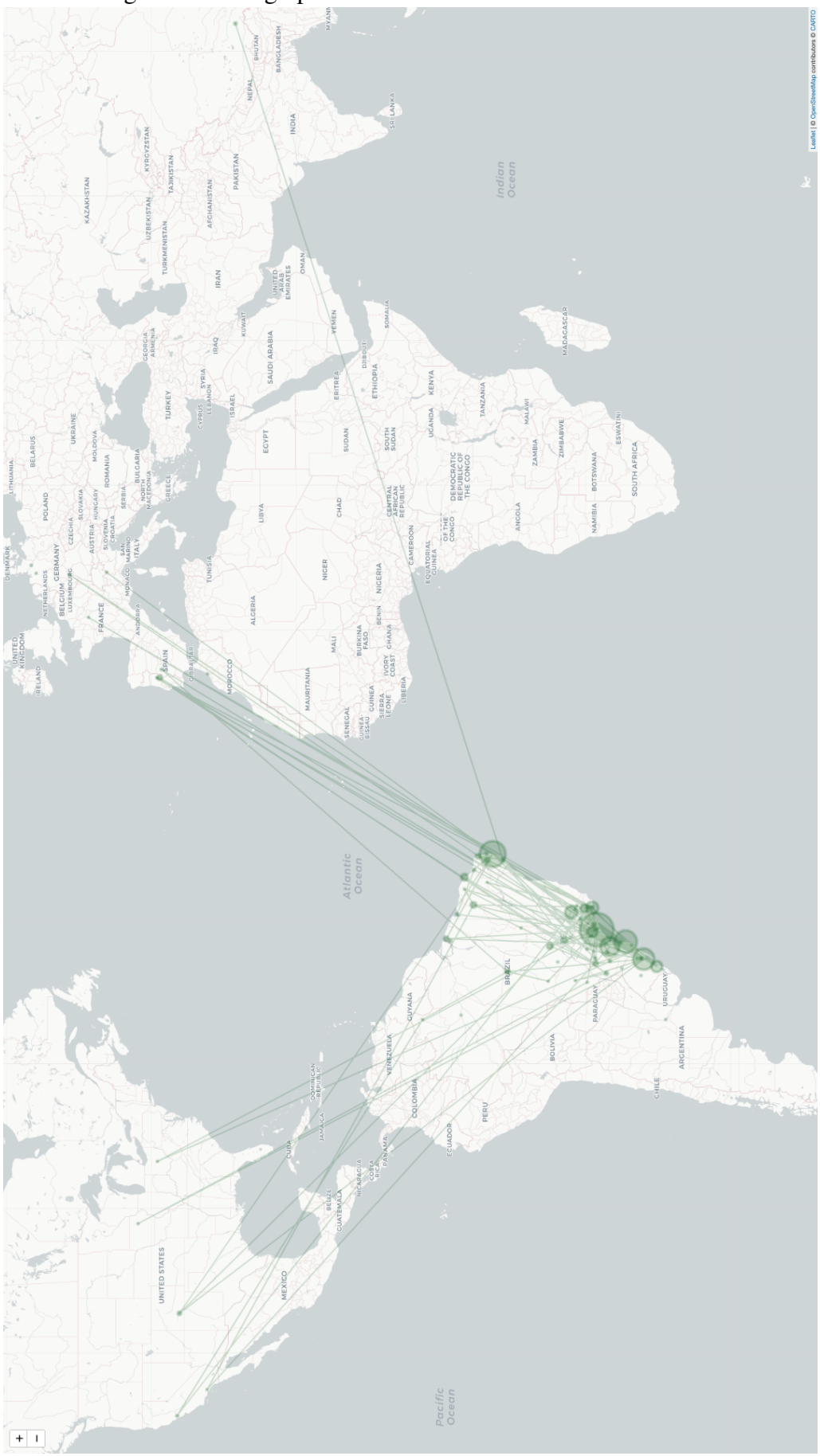
Author	*Institution*	*Frec*
FGF	INST015 - Brazil	4
CDMB	INST005 - Brazil	3
MN	INST025 - Iraq	3
NCSF	INST026 - Brazil	3
AKRD	INST001 - Brazil	2
APG	INST027 - Brazil	2
AARC	INST024 - Brazil	2
AC	INST028 - Brazil	2
BTB	INST009 - Brazil	2
BB	INST029 - India	2

Source: The Author

When we look at the frequency of authors published/rejection in the tables 4.15 and 4.16 the evidence that we have a wider range of submissions from different parts is reinforced, given us the notion of a more spread production across the map. And the most rejected authors indeed have lower numbers of rejections.

Now we will look at the author's distribution register at the publication period, to see across the map if our observations are true and which places RITA reached during those 15 years.

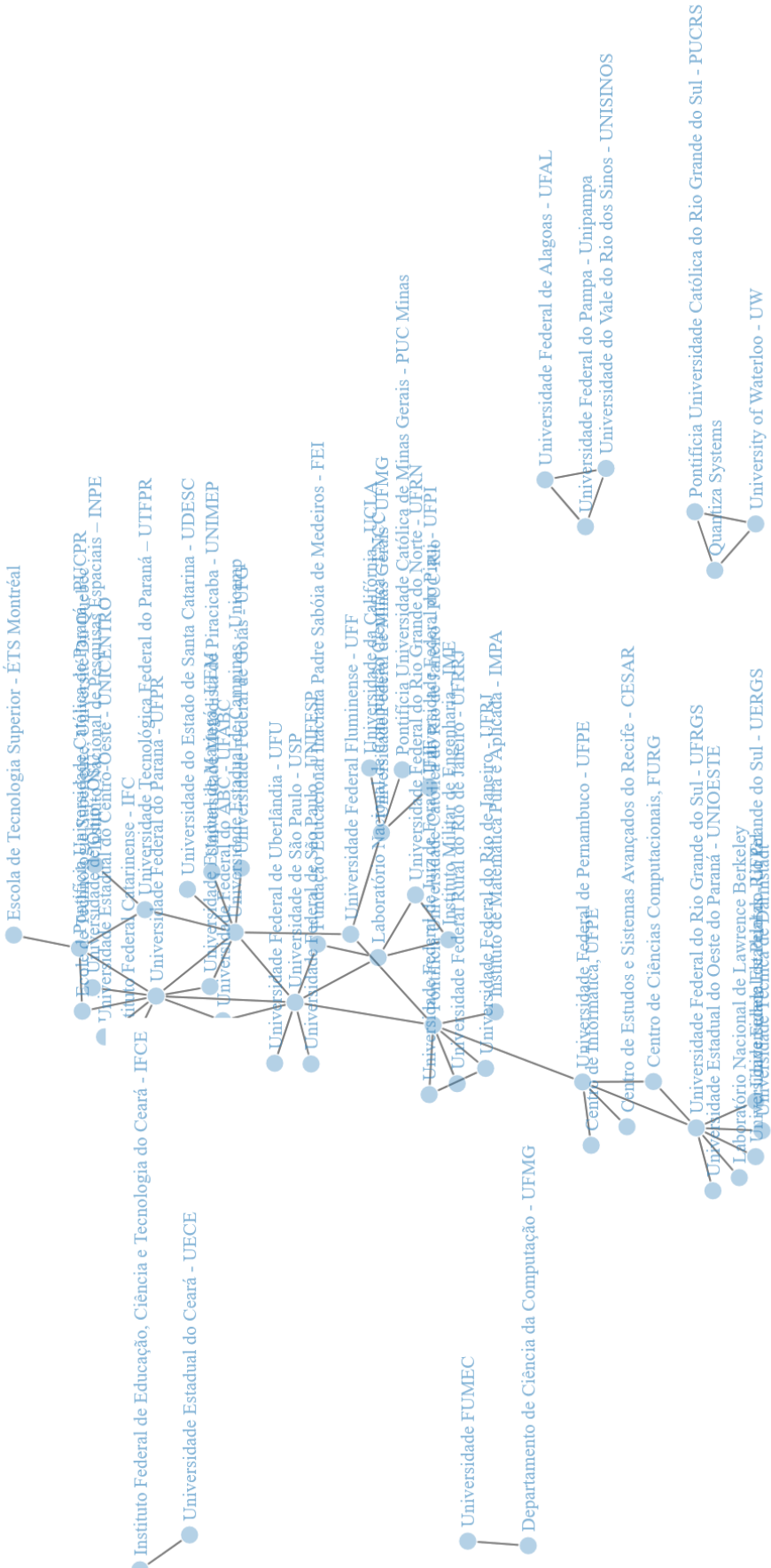
Figure 4.6: Geographic Location Of Authors With Published Articles



Source: The Author

The figure 4.6 confirms the claims made made in section 4.2 , having a wider distribution and confirms that RITA reaches more audiences outside Brazil. Its important to a journal that aims to reach higher classifications that the scope of audience can't be limited by local public. Despite not have a huge presence in external countries, this is an door opened for the journal expansion by gathering submissions from different parts of the world. Now that we have a notion of author distribution, we want to check the connections made between institutions based on the collaboration between authors on published studies at RITA.

Figure 4.7: Institutions Connections Registered Based on Co-Authorship



Source: The Author

An interesting observation is the collaboration between institutions becomes interconnected over time, giving more opportunities to see international collaborations published in the journal.

An analysis of the submissions to RITA reveals that our publication has reached a diverse array of universities across the country, as well as institutions abroad. The distribution of authors indicates a decentralized pattern of contributions, with submissions originating from various locations rather than being concentrated around a single author or institution. This trend is further highlighted when we examine the geographic distribution of authors, showcasing the wide reach of RITA's publications (see Figure 4.6).

Although the studies come from numerous different institutions, there is a noticeable level of interconnection when we look at collaborations between them (see Figure 4.7). This suggests that RITA has fostered a strong community of contributors who participate actively to its publications, generating international research and extending the journal's influence to a broader audience.

4.5 What are the key differences between accepted and rejected articles regarding keywords, topics, and other relevant metrics?

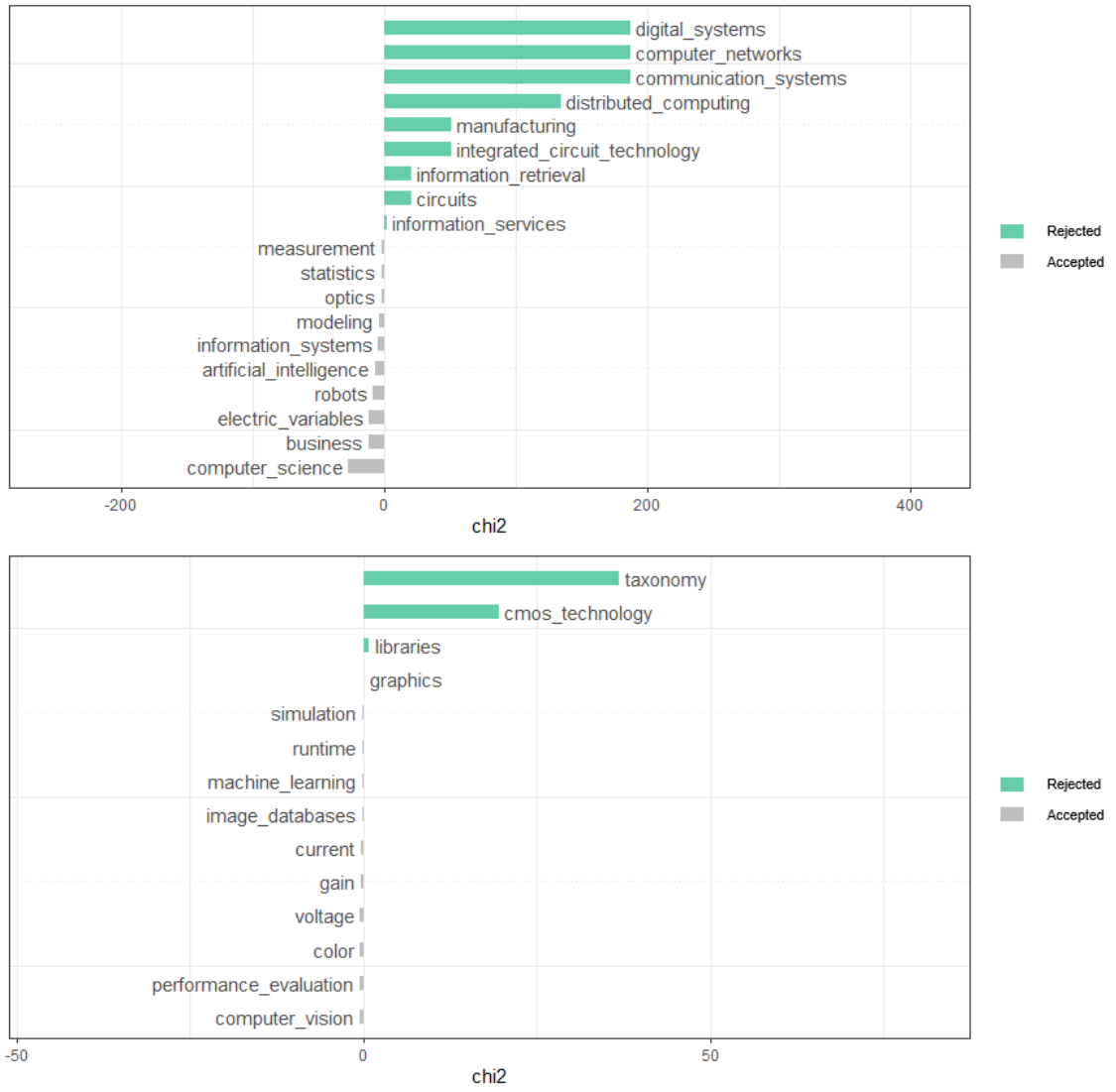
Our idea is to explore a little further this process by looking into the subjects brought from both accepted and rejected articles and comparing them to understand their characteristics. The idea follows the same principles presented in section 4.1, but instead of comparing the keywords of areas and subareas found in the selected years in contrast with previous years, we want to compare keywords found in abstracts of rejected articles and make a contrast test against keywords accepted in the same period.

4.5.1 Comparison of Accepted and Rejected Articles

Our last question is a form for us to take advantage of having in hand data related to rejected articles, to see if the keywords extracted from the abstracts of these articles are really that different in the thematic aspect. In the figures 4.8, 4.9, 4.10 and 4.11 we have our results when we crossed the rejected themes with the accepted ones.

We used snapshots from 2008, 2013, 2018, and 2023, where the target group represents the rejected keywords and the reference stands for accepted ones.

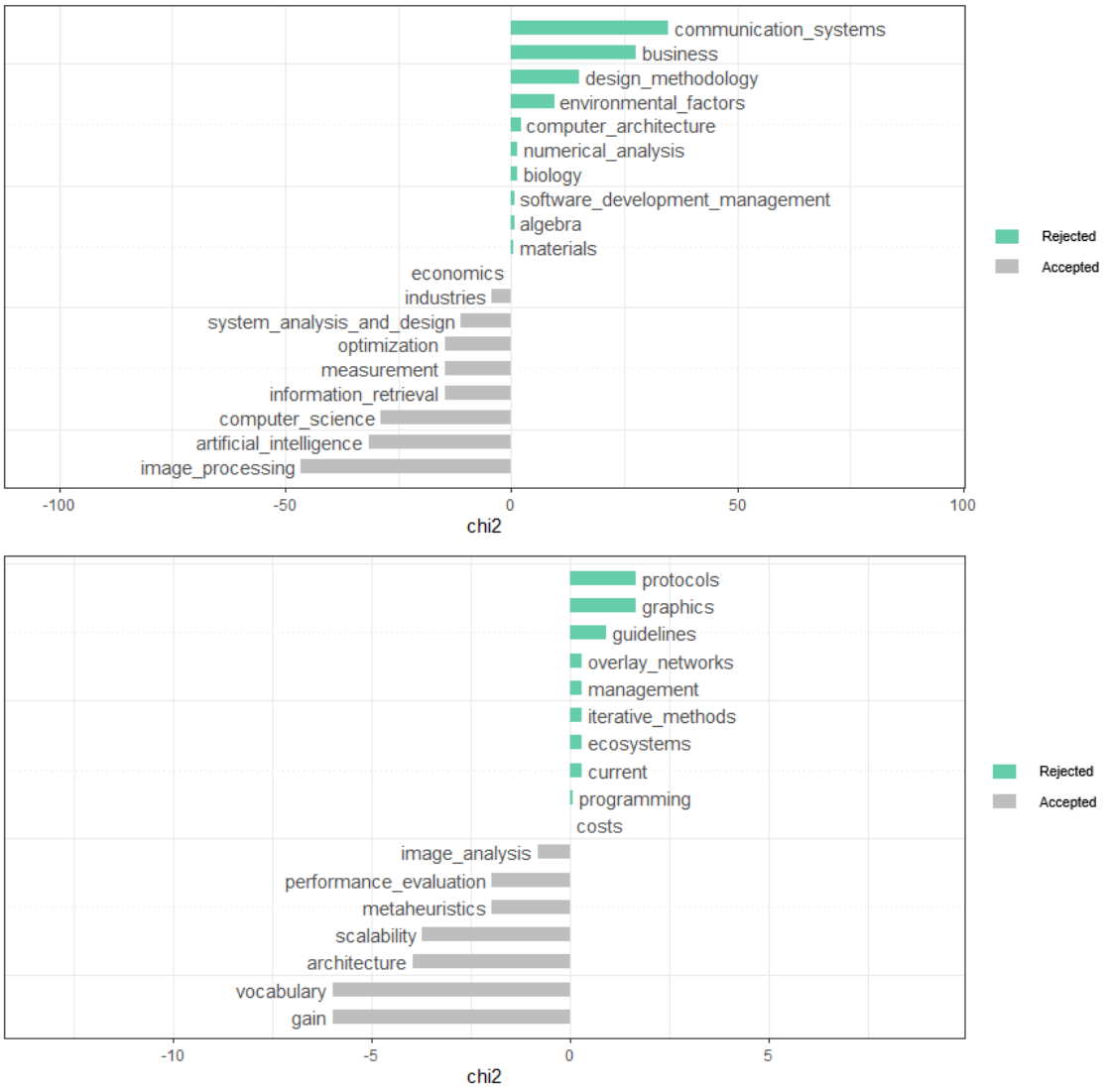
Figure 4.8: Comparison Of Rejected vs Accepted Areas and Subareas - 2008



Source: The Author

In the figure 4.8, the subjects chosen by the rejected articles are aligned with the topics of accepted articles. However, when we look at the subareas, they seem to be more concentrated than the accepted ones.

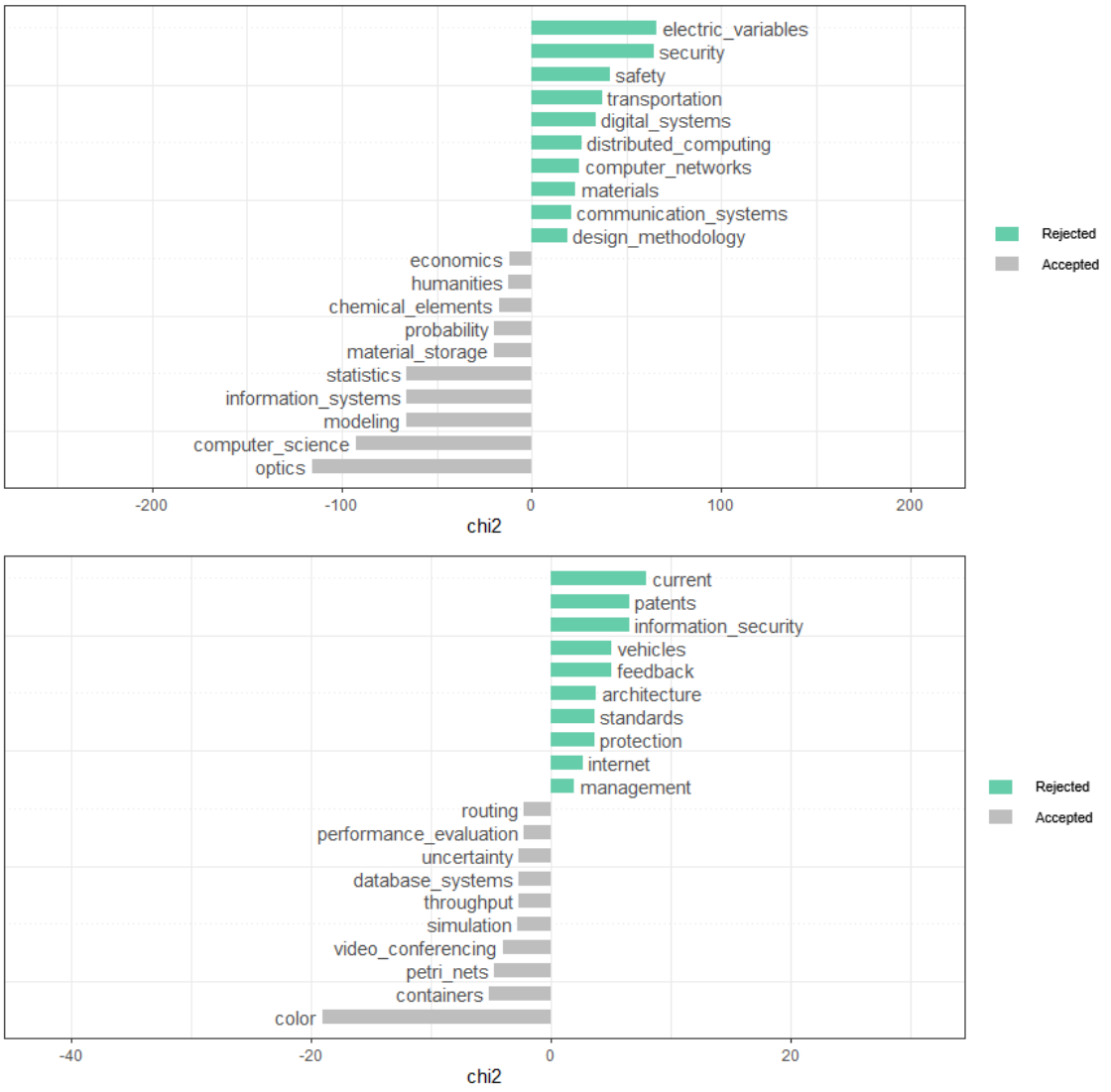
Figure 4.9: Comparison Of Rejected vs Accepted Areas and Subareas - 2013



Source: The Author

In the figure 4.9 the rejected keywords seem to take a more distance from the accepted ones, but on the other hand, the subareas of rejected topics have more variance. Another aspect noticed is the fact that business/management topics have made a strong presence here.

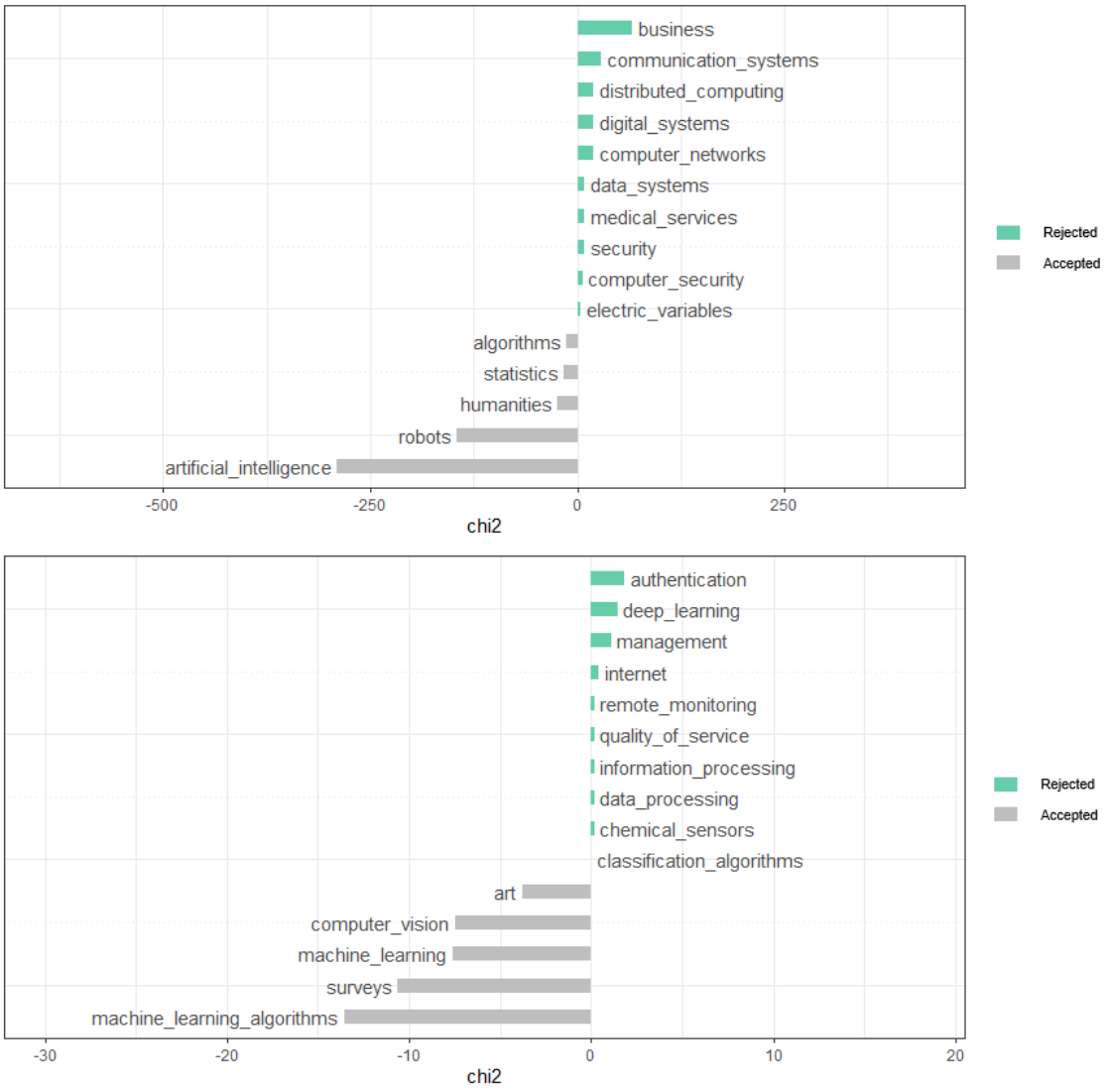
Figure 4.10: Comparison Of Rejected vs Accepted Areas and Subareas - 2018



Source: The Author

In figure 4.10 the rejected areas have more synergy with the accepted, both focused on computer engineering, but the accepted articles seem to be more focused on the communication studies while the rejected ones seem to be more divided with communication and information security, which indicates that the range of research of this year is more broaded in that sense.

Figure 4.11: Comparison Of Rejected vs Accepted Areas and Subareas - 2023



Source: The Author

In figure 4.11, the subjects seem to be more distributed, but even in these circumstances, we can see the presence of AI terms. Still, we can see business topics again, which is an interesting fact since business areas in computer science is very eager to invest in AI as the next big step of information technology services.

Despite noticing some similarities, we didn't find any remarkable evidence that points out a relevant characteristic that enlightens us on the sheer volume of rejected articles in RITA.

5 CONCLUSION AND AREAS FOR FUTURE EXPLORATION

In 2024, RITA will celebrate its 35th anniversary, and we aim to mark this milestone by exploring its 15-year history in digital format to gain insights into the evolution of our local scientific journal and identify patterns in the results. We chose bibliometric analysis to investigate RITA, and our findings, along with observed opportunities for future research, are detailed in this section.

5.0.1 Conclusion

The contribution of this study is to examine how RITA's productivity has evolved while also analyzing the information gathered during this process. The study delves deeper by exploring characteristics found in rejected articles, seeking insights that could reveal more facets of publication trends, particularly those observed in the rejected submissions.

Our first research question (RQ1) addressed the volume of submissions and the themes found in the journal. The data indicates that the claims made by Oliveira and Lamb (2004) are substantiated; however, despite a higher volume of submissions, fewer articles are being published. The themes found in RITA from 2008 to 2023 are largely aligned with the trending topics of each year, which is a positive sign, as it suggests that RITA has remained in sync with the academic trends within the field of computer science.

For our second research question (RQ2), we examined different indexers for each year to assess the journal's growth. The key observations are summarized below:

- The growth rate of publications has declined over the past four years.
- The average number of pages per article has decreased.
- There is an average of three authors per article.
- There has been a decline in metric views from 2017 to 2023.

Our third research question (RQ3) explores the topics that users search for in the journal by analyzing view counts to identify emerging trends in selected years and examining the most viewed articles in RITA's history. The data reveals that the most searched topics during the observed periods were AI, Computer Science, Economics,

and Robotics, while the most popular themes were recorded in 2008 (AI, Electric Variables, Robots, and Design Methodology).

For our fourth research question (RQ4), we analyzed the frequency of publications by the most prolific authors and institutions to understand the network RITA fosters between institutions and the global diversity of its contributors. The results show that RITA has received publications from a wide range of institutions, with the majority coming from Brazil and a significant number of submissions from around the world.

In the final research question (RQ5), we evaluated the topics and subtopics presented in rejected articles, comparing them with those in accepted articles to determine whether these terms were relevant when contrasted with the accepted submissions. The conclusion is that in certain periods, the topics in accepted articles resonated with those in rejected articles, likely due to the popularity of specific research trends.

RITA is currently facing a period of stagnation. Some of the reasons for RITA's decline can be attributed to the competition with more prestigious journals, limited resources, and a small editorial board. Our primary audience consists of newcomers to computer science, but to attract more readers, we need to gain a foothold among the more established segments of the academic community. One potential solution is to attract more renowned authors to publish in RITA, thereby increasing its visibility and appeal. Another approach could be to offer incentives to researchers who publish their work in RITA, making it a more attractive option. A final viable strategy is to narrow the scope of RITA's focus within the field of computer science, aiming to capture a niche audience and improve the conversion rate of high-quality published articles.

There's lots of challenges for a regional publication rise in the CAPES rankings under the current evaluation method for researchers, which authors give incentives to publish in well-ranked journals while new journals needs to capture quality studies to receive better Qualis classification. New journals must compete directly with well-established publications both domestically and internationally, and researchers will naturally prioritize submitting their work to higher-ranked journals. Despite its significance to the local computer science community in Brazil, RITA is going through a difficult period and must take decisive action to reinforce its position as a valuable part of history and to prevent its decline.

5.0.2 Areas for Future Exploration

For the last subject of this section, the idea is to pinpoint topics for future works because many insights appeared during the execution of this work that we couldn't address due to time constraints. This list of suggestions for future works is as follows:

1. This experiment lacks clustering analysis, which would be a great topic to improve even more, granting a richer analysis. Using clustering analysis is a great way to convey visual information and notice more emergent characteristics (DONTHU et al. 2021a).

2. Create a tool that consumes results generated from this pipeline and allows users to search metrics related to authors and institutions.

3. Explore more in-depth rejected articles to have more insights into the rejection aspect of RITA. Spatial analysis to understand where these studies came from and observe the interaction between the rejected articles can be meaningful in finding potential sources of low-quality studies and patterns between institutions. The exploration of rejected data is a niche not well explored in academic research and can be an opportunity to evolve this powerful toolbox.

REFERENCES

AGUINIS, H. et al. Debunking myths and urban legends about meta-analysis. **Organizational Research Methods**, v. 14, n. 2, p. 306–331, 2011. Available from Internet: <https://doi.org/10.1177/1094428110375720>.

ALSOLBI, I. et al. Different approaches of bibliometric analysis for data analytics applications in non-profit organisations. v. 2, n. 3, p. 90–104, 2022. ISSN 2767-6595 (Online). Publisher: OAE Publishing Inc. Available from Internet: <https://www.oaepublish.com/articles/jsegc.2022.09>.

AROMATARIS, E.; PEARSON, A. The systematic review: an overview. v. 114, n. 3, p. 53–58, 2014. ISSN 1538-7488.

BARIK, N.; JENA, P. Bibliometric portrait of select open access journals in the field of library and information science: A scopus based analysis. v. 2019, 2019.

DAS, S.; KAUR, K.; DR, M. K. V. Publication and collaboration pattern of college and research libraries journal during 2009-2018: A scientometric analysis. 2021.

DONTHU, N. et al. How to conduct a bibliometric analysis: An overview and guidelines. v. 133, p. 285–296, 2021. ISSN 0148-2963. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0148296321003155>.

DONTHU, N. et al. Forty years of the *International Journal of Information Management*: A bibliometric analysis. v. 57, p. 102307, 2021. ISSN 0268-4012. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0268401220315061>.

EDER, A.; FRINGS, C. What makes a quality journal? v. 65, p. 257–262, 2018.

ELLEGAARD, O.; WALLIN, J. A. The bibliometric analysis of scholarly production: How great is the impact? v. 105, n. 3, p. 1809–1831, 2015. ISSN 1588-2861. Available from Internet: <https://doi.org/10.1007/s11192-015-1645-z>.

GRAMES, E. M. et al. An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. v. 10, n. 10, p. 1645–1654, 2019. ISSN 2041-210X. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13268. Available from Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13268>.

GREENWOOD, D. C. Reliability of journal impact factor rankings. v. 7, p. 48, 2007. ISSN 1471-2288. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2206035/>.

GUPTA, S.; HASAN, N. Scientometric analysis of metamorphosis: A journal of management research. v. 38, p. 254–258, 2018.

HIRSCH, J. E. An index to quantify an individual's scientific research output. v. 102, n. 46, p. 16569–16572, 2005. ISSN 0027-8424. Place: United States Publisher: National Academy of Sciences.

JAMIE FOSTER SCIENCE. **The Rise Of Data Science: Why This Career Field Keeps Growing - Jamie Foster Science**. 2023. Section: FAQs. Available from Internet: <https://www.jamiefosterscience.com/why-is-data-science-a-growing-career-field/>.

KULKANJANAPIBAN, P.; SILWATTANANUSARN, T. Bibliometric analysis of publications in the scopus database: A study at prince of songkla university (PSU) during 1978-2021. 2021.

KUMAR, A.; AHMED, S. Challenges faced in the peer review system in open access journals. v. 13, p. 1–3, 2022.

KUMAR, R.; KALIYAPERUMAL, K. Scientometric analysis of global publication output in mobile technology. v. 35, p. 287–292, 2015.

LAMB, L. C.; OLIVEIRA, J. P. M. d.; GRANVILLE, L. Z. Editorial. v. 14, n. 2, p. 7–8, 2007. ISSN 2175-2745. Number: 2. Available from Internet: <https://seer.ufrgs.br/index.php/rita/article/view/rita_v14_n2_p7-8>.

LAMB, L. C.; OLIVEIRA, J. P. M. d.; GRANVILLE, L. Z. Editorial: 20 anos da revista de informática teórica e aplicada. v. 16, n. 1, p. 07–08, 2010. ISSN 2175-2745. Number: 1. Available from Internet: <https://seer.ufrgs.br/index.php/rita/article/view/rita_v16_n1_p07>.

MINGERS, J.; LEYDESDORFF, L. A review of theory and practice in scientometrics. http://authors.elsevier.com/sd/article/S037722171500274X, 2015.

OLIVEIRA, J. P. M. d.; LAMB, L. C. A revista de informática teórica e aplicada e as tecnologias de informação. v. 11, n. 2, p. 5–8, 2004. ISSN 2175-2745. Number: 2. Available from Internet: <https://seer.ufrgs.br/index.php/rita/article/view/rita_v11_n2_p5-8>.

PACHER, A. Who are the 100 largest scientific publishers by journal count? a webscraping approach. v. 78, p. 450–463, 2022.

RATHIKA, N.; THANUSKODI, S. Studies on relative growth rate and doubling time of publications productivity of nuclear medicine research. p. 198–211, 2021. ISSN 2456-9119. Available from Internet: <https://journaljpri.com/index.php/JPRI/article/view/2561>.

SAGE PUBLICATIONS. **Understanding Journal Metrics**. 2023. Available from Internet: <https://us.sagepub.com/en-us/sam/understanding-journal-metrics>.

SARKER, I. H. et al. Mobile data science and intelligent apps: Concepts, AI-based modeling and research directions. v. 26, n. 1, p. 285–303, 2021. ISSN 1572-8153. Available from Internet: <https://doi.org/10.1007/s11036-020-01650-z>.

SAVANUR, K.; SRIKANTH, R. Modified collaborative coefficient: a new measure for quantifying the degree of research collaboration. v. 84, n. 2, p. 365–371, 2010. ISSN 1588-2861. Available from Internet: <https://doi.org/10.1007/s11192-009-0100-4>.

SCHRöER, C.; KRUSE, F.; GóMEZ, J. M. A systematic literature review on applying CRISP-DM process model. v. 181, p. 526–534, 2021. ISSN 1877-0509. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S1877050921002416>.

SUBER, P. 6-7. ten challenges for open-access journals. 2016. ISBN: 9780262329552. Available from Internet: <https://knowledgeunbound.mitpress.mit.edu/pub/qfngzqxx/release/1>.

VERMA, S.; GUSTAFSSON, A. Investigating the emerging COVID-19 research trends in the field of business and management: A bibliometric analysis approach. v. 118, p. 253–261, 2020. ISSN 0148-2963. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S014829632030432X>.

VILLALOBOS-CID, M. **RPubs - Publish Document**. 2022. Available from Internet: <https://rpubs.com/mvillalobos/Rita>.

ÖZTüRK, O.; KOCAMAN, R.; KANBACH, D. K. How to design bibliometric research: an overview and a framework proposal. 2024. ISSN 1863-6691. Available from Internet: <https://doi.org/10.1007/s11846-024-00738-0>.