# Artificial Intelligence and Data Science Practices in Scientific Development

**Volume 8**

**Editor:** Alexandre Ribas Semeler

anis

advanced notes in information science

pro-metrics

# Artificial Intelligence and Data Science Practices in Scientific Development

**Volume 8** | ISSN: 2709-7587

**Editor:** Alexandre Ribas Semeler

anis advanced notes in information science

pro-metrics

**VOLUME EDITOR**

**Alexandre Ribas Semeler,** *Universidade Federal do Rio Grande do Sul, Brazil*

**COVER DESIGN**

**Chico Maciel,** *with background photography by Umberto (from Unsplash.com)*

**INTERIOR DESIGN**

**Chico Maciel,** *São Paulo, Brazil.*

**FIRST PUBLISHED 2025 BY PRO-METRICS OÜ**

Sakala 7-2, 10141 Tallinn, Estonia
https://pro-metrics.org

**COPYRIGHT INFORMATION**

© 2025 Pro-Metrics OÜ. The publisher owns the copyright over the entire book.
© 2025 The authors. Authors retain copyright over their chapters, licensed under a Creative Commons license CC BY-NC 4.0.

**PUBLICATION DETAILS**

ISBN: 978-9916-9331-3-8
ISBN: 978-9916-9331-4-5 (PDF)
ISBN: 978-9916-9331-5-2 (EPUB)
ISSN: 2709-7587
EISSN: 2709-3212
DOI: 10.47909/10.47909/978-9916-9331-4-5.13

**PRO-METRICS OÜ**

For inquiries related to this work's copyright and reproduction rights, please get in touch with us at editorial@pro-metrics.org

**ABOUT THE SERIES**

The Advanced Notes in Information Science (ANIS) book series publishes conference proceedings, monographs, and thematic volumes that explore the nexus of information, communication, and computer sciences. The ANIS series considers research works covering a range of topics, including but not limited to information retrieval, information systems, information architecture, information behavior, digital libraries, information literacy, information management, data management, library studies, user experience design, knowledge management, sociology of information, science communication, mass communication, organizational communication, and others. The series is intended to serve as a platform for students, researchers, and practitioners from the public or private sectors.

**SERIES EDITOR**

**Adilson Luiz Pinto,** *Federal University of Santa Catarina, Brazil*

**EDITORIAL BOARD**

**Rafael Capurro,** *Stuttgart Media University, Germany*

**Rosa Lidia Vega-Almeida,** *Biocubafarma, Cuba*

**Andrea Hrčková,** *Slovak Technical University, Slovakia*

**David Caldevilla Domínguez,** *Complutense University of Madrid, Spain*

**Jesús Lau,** *Universidad Veracruzana, Mexico*

**Carlos Alberto Ávila Araújo,** *Federal University of Minas Gerais, Brasil*

**Jela Steinerová,** *Comenius University in Bratislava, Slovakia*

**Almudena Barrientos Báez,** *University of La Rioja, Spain*

## TABLE OF CONTENTS

## PREFACE

This book constitutes a comprehensive exploration of two pivotal domains that are indispensable for comprehending the contemporary challenges confronting science and society: artificial intelligence (AI) and data science. These disciplines profoundly influence the generation of knowledge, the development of innovations, and the resolution of complex problems in various domains of social, economic, political, and cultural life. Entitled *Artificial Intelligence and Data Science Practices in Scientific Development*, this book compiles studies and applications from research groups collaborating at the intersection of technology, information, and scientific innovation and cultural areas.

These groups include *Technopoetics, Digital Art, and Neuroaesthetics*. *Creativity* investigates the physiological foundations of aesthetic experience, exploring topics such as the philosophy of technology, communication, digital art, and the impacts of AI on the visual arts. The *Data Science and Engineering Laboratory* conducts telehealth research. The *LEMME Lab's* research in AI focuses on its applications in services and products. The *Metric Studies in Data Librarianship and Geosciences* research endeavors encompass the domains of computational science and metrification. The *Sphere Information Ecosystem in Science, Technology, Innovation, and Sustainability* has three primary initiatives: data engineering, interoperability, and research information systems.

This collection comprises nine chapters, each one addressing distinct aspects of the application of computational intelligence and data analysis in fields such as health, the arts, political science, bibliometrics, scientific communication, and knowledge management. The outcome of this convergence is presented in these chapters. Despite the heterogeneity in the objects and methodologies of the chapters, they are unified by a shared axis: the pursuit of technological solutions that are firmly rooted in scientific foundations. This pursuit entails a meticulous examination of the social, ethical, and epistemological ramifications of automation, algorithmic analysis, and extensive data processing.

The first chapter, entitled *Art, Technology, and Creative Processes: A New Paradigm for Artistic Production*, offers a thought-provoking reflection on the impact of AI and neuroscience on creative processes in contemporary art. The central question guiding this study is as follows: How are algorithms reshaping the artist and

creativity in the 21st century? The integration of foundations from art theory, neuroimaging, and computational intelligence is demonstrated in the text, which reveals how creative processes can be simulated and expanded by algorithmic models capable of emulating human cognitive patterns. The chapter proposes a concept that extends beyond the realm of mere automation in aesthetic production, namely that of computational mannerism. This novel form of creation is defined by the human–machine interface, which serves as the medium through which novel artistic expressions emerge, thereby challenging the conventional boundaries of authorship, intuition, and originality. By problematizing the role of the artist in the age of AI, the chapter opens the book with a bold, transdisciplinary approach that challenges the traditional paradigms of aesthetic creation.

The second chapter, entitled *Obstetric Decision-Support System: An Informational Model for Maternal Autonomy Towards the Agenda 2030 Health Goals*, sets out an informational model for maternal autonomy in relation to the United Nations' Agenda 2030 health goals. The model focuses on the interface between information, health, and reproductive autonomy. The study recommends a model of an informational system designed to support decision-making by pregnant women, in light of the Sustainable Development Goals of the 2030 Agenda, especially SDG 3. The authors have developed an informational architecture of dynamic and static layers, integrating World Health Organization guidelines, lived experiences, and personalization algorithms. The system has been developed for the purpose of mitigating information asymmetries in obstetric contexts, with the objective of promoting user empowerment without compromising clinical safety. This chapter is noteworthy for its integration of scientific evidence, principles of humanized care, and requirements engineering methodologies, offering a replicable model tailored to the needs of vulnerable populations.

The third chapter, entitled *Data Provenance and Blockchain: An Approach in the Context of Health Information Systems*, discusses the application of blockchain technology to data traceability and integrity in health information systems. The objective of this study is to investigate the potential contributions of data provenance and the immutable attributes of blockchain to the security, interoperability, and reliability of health data. A comprehensive review of the extant literature and a qualitative analysis were conducted,

with a specific emphasis on electronic health records (EHR) and personal health records (PHR). By cross-referencing international interoperability standards with the technical characteristics of distributed systems, the authors construct a critical and propositional analysis of the limits and potential of using blockchain in this strategic sector. The chapter proposes methodologies for the development of data ecosystems that are characterized by enhanced transparency, auditability, and patient-centeredness.

In the fourth chapter, entitled *Structuring a Data Lake for the Management of Scientific Information in Brazil*, the initial steps in constructing a data lake designed to organize scientific information within the Brazilian research information system (BrCris) are outlined. The authors delineate a meticulous technical process for the collection, transformation, indexing, and visualization of scientific data from platforms such as OpenAlex and DOAJ. Utilizing techniques such as author disambiguation, cross-referencing data by DOI, and journal stratification, the chapter demonstrates the potential for the organization of substantial scientific data into structures that facilitate robust inferences, institutional intelligence, and strategic planning of national science. This case exemplifies the implementation of data engineering methodologies within the context of public science administration.

In the fifth chapter, entitled *Dialogic Bridges: Voices between Ideological Frontiers*, the focus is shifted to the political sphere, and a methodology is proposed based on network analysis to identify parliamentarians with greater potential for articulation between ideological blocs. Utilizing metrics such as bridge coefficient, betweenness centrality, and bridge centrality, the study suggests a replicable methodology for identifying political mediation agents in multiparty and fragmented contexts. By analyzing voting patterns and agreements in legislative houses, the chapter contributes to studies on governability, coalition building, and the mitigation of extreme polarizations. In this context, AI is employed as a tool for the structural analysis of complex social phenomena.

The sixth chapter, entitled *Analysis of Patent Production in Brazil: A Perspective from the Lattes Platform,* undertakes an analysis of the production of patents in the country, utilizing data extracted from the Lattes platform. The analysis correlates the patent production with the level of training of researchers. Through the development and application of extraction and validation algorithms, the authors identify patterns of technical production

over time and delineate an institutional panorama of the culture of innovation in Brazil. The findings indicate a notable concentration of patents among researchers who hold a doctoral degree, thereby corroborating the hypothesis that heightened inventive capacity is closely associated with advanced educational training and academic integration. The integration of scientific and technological metrics with systematic data analysis offers a valuable contribution to the development of public innovation policies.

In the seventh chapter, entitled *A Framework for Collecting, Processing, and Analyzing Scientific Data on Social Media*, the authors propose the Social4Science platform, which is designed for the collection and analysis of social data associated with the dissemination of scientific publications. The primary focus of the Social4Science platform is the YouTube platform. This study explores the emerging field of altmetrics, investigating how scientific research is received, commented on, and shared on social networks. This investigation aims to broaden the understanding of contemporary flows of knowledge circulation. The platform enables the correlation of social data with scientific metadata, facilitating analysis of trends, patterns of engagement, and the social impact of science. The chapter presents an innovative contribution to the field of science communication by suggesting tools that facilitate greater public engagement with science in digital environments.

The eighth chapter, entitled *Design Without Data? A Study of Methodological Transparency in Contemporary Design Science*, problematizes the scarcity of methodological transparency in design publications. A comprehensive analysis of over 7,500 articles from specialized journals reveals that only a limited number of these articles explicitly articulate their methodological approaches. Utilizing OpenAlex for metadata collection and ChatGPT-4o for automated classification of abstracts, the authors identify a predominance of speculative, conceptual, or practice-based research, devoid of declared methodological rigor. The chapter puts forth the argument that standardization and qualification of methods in design science are imperative for enhancing its credibility and facilitating interdisciplinary integration.

Finally, the ninth chapter, entitled *Bibliographic Analysis of Scientific Literature on Health Knowledge Management*, carries out a bibliometric analysis of scientific production on Health Knowledge Management (HKM) between 1990 and 2023, based

on data from Web of Science and PubMed. Based on the co-occurrence of terms and the mapping of authors and institutional collaborations, the study identifies four central thematic axes: the impacts of COVID-19 on health information management, strategies for improving the performance of health systems, challenges related to electronic medical records, and advances in big data and information technologies. The chapter reveals the centrality of the United States in academic production on the subject and highlights the need for investment in technological infrastructure and international collaboration to promote more effective health systems integrated with knowledge management.

This book is, therefore, an interdisciplinary mosaic that combines theory and practice, critical analysis and technological development, and experimentation and reflection. The chapters herein demonstrate the diversity of applications of AI and data science in scientific development. Moreover, they highlight the importance of approaches informed by ethical values, social commitment, and epistemic responsibility.

The integration of diverse disciplinary domains in this collection underscores the imperative for contemplating technological solutions that are not divorced from human needs, public policies, cognitive justice, and the democratization of knowledge. I would like to express my profound gratitude to the authors who have contributed to the creation of this book, dedicating their time and knowledge to writing consistent, innovative, and relevant chapters.

Additionally, gratitude is extended to the institutions that promote research and the generation of interdisciplinary knowledge, particularly those that have directly or indirectly contributed to the realization of this work. It is our hope that perusing this book will inspire reflection, provoke questions, and stimulate the creation of new investigative paths at the intersection of technology, data, and knowledge.

**Dr. Alexandre Ribas Semeler**
*Universidade Federal do Rio Grande do Sul*

CHAPTER 1

# ART, TECHNOLOGY, AND CREATIVE PROCESSES: A NEW PARADIGM FOR ARTISTIC PRODUCTION

**Alberto Marinho Ribas Semeler**
PPGAV, *Federal University of Rio Grande do Sul, Brazil.*
ORCID: *https://orcid.org/0000-0003-3380-9781*

**Alexandre Ribas Semeler**
*Geosciences Institute, Federal University of Rio Grande do Sul, Brazil.*
**Email:** *alexandre.semeler@ufrgs.br*
ORCID: *https://orcid.org/0000-0002-8036-4271*

## ABSTRACT

This study explored the evolving relationship between contemporary art, artificial intelligence (AI), and neuroscience, challenging the anthropocentric notion of artistic creation as uniquely human. The research question to be analyzed was as follows: "How are algorithms reshaping the artist and creativity in the 21st century?" To address this question, the integration of concepts from art theory, neuroscience, and AI was considered. This examination explored the manner in which neuroimaging technologies and biometric algorithms were reshaping our understanding of creativity. The study examined the impact of scientific progress on artistic expression across different eras, ranging from the advent of psychoanalysis to the emergence of computer technologies. It demonstrated how neuroscience was facilitating our understanding of the brain processes underlying creativity, including the neurotransmitters and cortical regions implicated in artistic processes. Empirical analyses were supported by neuroimaging

studies that established a correlation between brain activity and aesthetic experiences, as well as algorithmic simulations that simulated artistic cognition. Recent findings indicated an increasing role for AI in artistic production, with the technology emulating the brain's creative processes. The neurotransmitters dopamine and oxytocin were demonstrated to influence artistic motivation and pleasure. Furthermore, neuroimaging studies showed that creative activities resulted in the activation of regions such as the limbic system and the prefrontal cortex. The extension of these processes enabled algorithmic models to generate artworks that defied conventional art definitions. The investigation introduced computational mannerism, a concept in which digital interfaces and machine learning expanded artistic potential by reflecting human cognitive patterns in real-time iterations. This suggested a fusion of human intuition and machine logic, thereby challenging the exclusivity of human creativity. This integration of neuroimaging data into algorithmic systems represented a paradigm shift, giving rise to a range of ethical and philosophical questions concerning authorship, creativity, and the artist's role in the digital age. As AI progresses, it became imperative to develop novel theoretical frameworks to comprehend its cultural and metaphysical influence on artistic expression.

**KEYWORDS**: assisted creation, computational mannerism, aesthetics of artificial intelligence, biological pathway in creative process, neuroimages

**HOW TO CITE**: Semeler, A. M. R., & Ribas Semeler, A. (2025). Art, technology, and creative processes: A new paradigm for artistic production. In A. Semeler (Ed.), Artificial Intelligence and Data Science Practices in Scientific Development, *Advanced Notes in Information Science, volume 8* (pp. 11 - 29). Pro-Metrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9331-4-5.52.

**COPYRIGHT**: © 2025 The author(s). This article is distributed under the terms of the CC BY-NC 4.0 license, which permits copying and redistribution of the material in any medium or format, adaptation, transformation, and building upon the material, provided that the license terms are followed.

# 1 INTRODUCTION

The utilization of novel technologies, whether digital or analog, by artists to modify computer systems, devices, and software programs, as well as to engage in collaborative endeavors with engineers and programmers, has been a persistent feature of artistic production. Historically, technology has been repurposed for creative purposes, despite its origins in entirely different applications. Given the paucity of software applications dedicated to artistic creation, software intended for commercial or industrial use is often adapted to align with the creative process. This technological adaptation is driven by the necessity to expand creative limits and innovate within the artistic field. The primary inquiry guiding this investigation is as follows: How are algorithms reshaping the artist and creativity in the 21st century? This study integrates concepts from art theory, neuroscience, and artificial intelligence (AI) to investigate the ways in which neuroimaging technologies and biometric algorithms are restructuring our comprehension of artistic expression and creativity. The objective of this study is to analyze the appropriation of digital and analog technologies by artists to expand creative possibilities through technological adaptation and to examine the manner in which the practice of repurposing software designed for nonartistic purposes contributes to the process of artistic innovation. Furthermore, the research examines the influence of AI and neuroscientific principles on contemporary art, highlighting how technological mediation reshapes creative processes and redefines artistic expression in the digital age. In the domain of design, certain corporations, such as Adobe, have developed dedicated programs that facilitate artistic production.

Nevertheless, the notion of creation as a genuinely innovative act—one that, as Boden (2007) elucidates, represents "something that no one else has ever done before"—exceeds mere digital manipulation. The author posits a dichotomy between two types of creation: historical creation, also referred to as H-creativity, and original creation, which is characterized as a catalyst for advancement. Alternatively, personal creation, or P-creativity, does not include the majority of people who produce only average ideas, already known by others, although they are new to the individual in question. The act of creation, therefore, must entail the establishment of a novel paradigm for the collective. The

essence of art is encapsulated in its capacity to act as a disruptive force that transcends the boundaries of technology. The correlation between art and science is evident. Art frequently borrows methods and tools from scientific advancements, while science frequently pursues innovation and creativity through artistic experimentation. The ensuing discourse is methodically structured into distinct sections, each addressing a distinct yet interconnected facet of the intricate relationship between AI, art, and technology within the ambit of creative processes. The initial section of the text examines how artists employ analog and digital technologies, emphasizing the importance of hacking and repurposing technology due to the paucity of specialized software designed for creative innovation. This necessity for adaptation underscores the artist's role as both creator and technological manipulator, propelling the boundaries of conventional tools to achieve novel forms of expression.

Subsequently, a historical analysis is conducted to examine the impact of scientific theories on artistic practices. Concepts such as psychoanalysis and early AI investigations have had a profound influence on the development of visual and computational technologies. This study will examine the evolution of technological devices in the context of 20th- and 21st-century art. It will demonstrate how works of art are inextricably linked to the technologies available in each period, including photochemical, electronic, computational, and digital technologies. This exploration of art historiography delineates the evolution of artistic media from the advent of photography and cinema (photochemical art) to digital art. According to Paul (2015), a theorist of technologies applied in the field of arts, she proposes the term "digital art" as an umbrella to describe the great proliferation of styles that emerge from new technological devices. With each technological debut, a novel style emerges, complicating its analysis within the framework of art theory. These technological shifts signify profound transformations in creative practices and aesthetic critique, thereby unveiling a growing divide between aesthetic theory and artistic execution. This discrepancy underscores the necessity for novel theoretical frameworks that address the integration of digital technologies in contemporary art.

In its contemporary analysis, the text addresses the role of AI in creative processes, focusing on the cognitive and neurobiological mechanisms that underlie artistic production. The following

analysis will examine the brain processes that occur during the creative process and will delineate the specific stages in which brain areas and neurotransmitters act. This model proposes a methodology for the construction of a creative computer. Indeed, the act of creation in art necessitates the engagement of cognitive and affective processes. This study examines the participation of neurotransmitters and specific brain regions in creativity, highlighting how computational models can simulate these biological processes to replicate or even enhance artistic expression. This intersection of neuroscience and AI suggests that creativity, traditionally viewed as an exclusive human trait, is increasingly accessible to algorithmic replication, thereby challenging long-held beliefs about the uniqueness of human artistic capability. The section dedicated to visual interfaces explores how these technologies mediate the relationship between the human mind and the digital world, creating new possibilities for aesthetic interaction and expression. Interfaces designed with algorithmic precision have been shown to facilitate immersive experiences that extend artistic perception beyond the physical limitations of traditional media. This technological mediation enhances the accessibility of art and redefines its experiential dimension, proposing a paradigm shift toward more interactive and responsive artistic creations.

The proposed methodology delineates a framework for digital creative processes that integrates concepts of art, neuroscience, and technology. The text places significant emphasis on the importance of algorithmic literacy, underscoring the necessity for creators to possess a comprehensive understanding of, and deliberate control over, digital interfaces and algorithms. This literacy is presented as essential for navigating the complexities of AI in artistic production, where the artist's role evolves from mere creator to orchestrator of digital processes. By means of this argumentative structure, the discourse explores how the algorithmic revolution has reshaped contemporary art. It challenges traditional notions of creativity and redefines the artist's role in a digitally mediated environment, marking the dawn of a new artistic era—one where human intuition and machine learning coalesce to push the boundaries of artistic expression.

## 2 LITERATURE REVIEW

The methodology for conducting a literature review is structured into distinct sections that trace the development of artistic methods in response to technological progress. The analysis initiates by examining the influence of psychoanalysis on artistic movements such as Surrealism and Dadaism. The study emphasizes the impact of delving into the subconscious on shaping creative approaches. The selection of Freud's theories is attributed to their groundbreaking proposition that an unconscious mind governs the mechanisms of consciousness. The concept of the "narcissistic wound," introduced by Freud, is also discussed as a metaphor shedding light on human apprehension toward the rise of AI. The narrative then charts the gradual assimilation of technology into art, starting from traditional photochemical techniques such as photography and cinema to the realm of electronic and computational art, culminating in the utilization of digital tools and AI. This progression underscores a widening gap between artistic practice and aesthetic theory, characterized by transformations in creative methodologies and the conceptual function of the artist. The incorporation of insights from neuroscience into the development of algorithmic and computational models is deemed increasingly essential. This integration draws on an understanding of brain functionality, specialized neural regions, and models that have influenced the advancement of AI. Since the inception of the field of cognitive computing, the human brain has served as a foundational reference point, influencing the design of both the hardware components and the software programming of artificial devices.

### *2.1 Historical approach: Historiography of art and technological devices*

Since the advent of the 20th century, with the emergence of psychoanalysis, the field of art has sought to explicate its own functioning, as evidenced by the emergence of Surrealism and Dadaism. Psychoanalysis is a theoretical framework that can be employed to elucidate certain creative strategies employed by artists. For instance, in developing his poetics, he employed psychoanalytic theory, particularly the concept of the unintentional,

thereby gaining access to the unconscious. It is important to acknowledge that Freud, a renowned neurologist, possessed a profound understanding of brain physiology when he developed psychoanalysis. This is the rationale behind its employment in this context. His theoretical contributions have exerted a profound influence on artistic expression and creative thinking throughout the 20th century. In addition to his theoretical work, he was also an expert in the field of brain mechanisms.

In contemporary discourse, psychoanalysis is often stigmatized as pseudoscience; nevertheless, its seminal discoveries concerning the unconscious continue to exert a substantial influence on the field of neuroscience. In his reflections on the emotional brain, neuroscientist Joseph LeDoux illuminates Freudian notions. Freud's concept of the unconscious as a storage space for conscious content is a seminal one. However, it is important to note that the unconscious is also a repository for thoughts and memories of fear and anxiety, which are stored and maintained in a way that is inaccessible to the subject. The cognitive unconscious refers to processes that activate functions that may or may not produce conscious content. When discussing processes of this nature, I opt for the term "nonconscious" to avoid confusion with the Freudian unconscious. The author proposes that emotions arise in three levels of conscious, nonconscious, and unconscious feelings. In contrast to genuine fear, anxiety is a construct of our psycho-corporeal response, stemming from adrenaline mechanisms and learned behaviors. The concept of the unconscious is theorized as a physiological response to various emotional stimuli experienced by the body. This cerebral metabolism, or nonconscious process, of anxiety and fear is present in all emotional reactions, including aesthetic reactions. In essence, he is refining certain tenets of Freudian theory. These reflections are instrumental in our proposal, as AI is capable of analyzing and precisely connecting with such visceral and nonconscious reactions. LeDoux is credited with the conceptualization of the emotional brain, and in this work, he reviews several positions on the mechanisms by which affective triggers are initiated within the cerebral amygdala, adrenaline, and noradrenaline (LeDoux, 2015).

The prevailing notion is that the fear circuit in the brain is responsible for the sensation of fear. When activated, this circuit instigates characteristic responses in humans, including paralysis, facial expressions, and alterations in body physiology. The

phenomenon of fear is frequently regarded as an intermediary between a perceived threat and the subsequent physiological and behavioral responses. Fear is a genuine phenomenon; certain factors contribute to the behavioral threat. In this study, we propose the concept of "like a key" as a theoretical framework to elucidate human responses to AI. The following inquiry is posited: what form does an imaginary fantasy assume in the aftermath of the loss of the anthropocentric protagonist of intelligence? In the historiography of art, our proposal follows a specific path that foresees demarcated phases in the use of techniques, information theory, cybernetics, and computational and digital technologies in the arts. In this article, we will employ the principles of aesthetic theory, tracing the development of this concept from its origins in the works of Kant to the present day. In their reflections on art and creativity, some authors eschew contemporary aesthetic theory. This phenomenon can be attributed to the inherent complexity and distribution of cognitive experiments. We do not seek to invalidate these authors; on the contrary, in our reflection, we will seek to add them to our proposition and reflection on art and creation (Vartanian et al., 2013). An illustration of this complexity is the pervasive and simplistic use of the term "art and technology." This term is employed to denote art that utilizes technologies of various types and eras.

The proposal delineates discrete periods characterized by distinct formal and evolutionary distinctions from the analog and hybrid technologies of the 20th century to the digital and intelligent technologies of the 21st century. When the term is employed in a general sense, it fails to acknowledge the poetic, formal, and creative potential inherent in each technological device. These structural differences are the result of the aforementioned factors. So, we consider photography and cinema (photochemical art); analog means of telecommunications such as mail, telephone, and radio (art and communication); television video (electronic art); computational technologies (computational art); and currently the digital computer, information and communication technologies, and AI (digital art). This process commenced at the onset of the 20th century and subsequently intensified, engendering profound transformations in the realm of art from the 1950s through the 1960s. With the cessation of the Second World War, the technology industry no longer enjoyed the level of funding that had previously been provided by wars

and their technological war experiments. This pivotal moment is concomitant with a gradual relinquishment of academic and conventional pretensions in art, which sought to uphold the constraints imposed by both art in relation to traditional techniques and their circumscribed domains, and aesthetics in relation to its ontological foundations (the thematic branch of philosophy that appreciates beauty).

These novel proposals engender profound transformations, which are not always comprehended or embraced by the artistic community. This issue is further complicated by the ongoing and intensifying controversy surrounding the crises in art and aesthetics. At the core of this controversy, which has been disseminated by certain postmodern theorists, all indications pointed to a purported dissolution of both fields: art and aesthetics. This controversy emerges, at least in part, from the pursuit of beauty and its reflection in the philosophical domain by an aesthetic theory. It is evident that art and creative practices are undergoing a paradigm shift, marked by a gradual deconstruction of their underlying principles. The incorporation of the abject and the unpleasant as values to be appreciated, as well as situations of total neutrality without a priori valuation of either one or the other, is a key tenet of the philosophy. The phobic's sole object is the abject. Therefore, with fear in parentheses, the discourse will appear sustainable only if it continuously confronts this otherness, a burden that is simultaneously repulsive and repelled, a deep memory that is inaccessible and intimate: the abject (Kristeva, 1982).

Consequently, the role of aesthetics and the functions of art diverge. To achieve this objective, it is imperative to prevent governments and large corporations from exerting control over AI. The artistic process necessitates the liberty to employ any form of information, irrespective of its aesthetic quality, neutrality, or the extent to which it may be regarded as aesthetically displeasing. The increasing use of technologies as artistic tools has led to a significant and ongoing division between artistic experience, art criticism, and aesthetic theory. It is imperative to note that, while these elements should maintain synchrony and congruence in their correspondence, they have, in fact, embarked on discordant trajectories. The dissonance between the theoretical corpus and artistic practice has engendered a paradox, which is arguably a primary factor in the persistent declaration of the "death" of art

in the 20th century and, more recently, the "death" of the artist in the 21st century with the emergence of creation by AI. In another vein, during the 20th century, the advent of computational technologies precipitated a systematic investigation into the human neural apparatus. This investigation was driven by the development of computers and AI. It also sought to elucidate the mechanisms underlying creativity and the cognitive processes involved.

Preliminary research suggests that the components of computational machinery are influenced by the activation of neurons in the human cortex. This trigger is employed in research as a multidisciplinary paradigm that shifts the manner in which computers operate. Consequently, it establishes a novel cultural perspective on the opportunities for the style of creation by AI. Our gaze is subject to and respects the rules of brain functioning. For instance, ultraviolet light is not visible to the human eye and has never been represented in any artistic medium. We are subject to the laws of the brain (Zeki, 1999). Consequently, the utilization of ultraviolet light in the creative process is only feasible through the implementation of specialized software that encodes the pulses of this light, which is imperceptible to the naked eye, without the use of sensors. This paradigm shift has been particularly evident in certain scientific disciplines, marking a transition from the humanities to the exact sciences. This multidisciplinary approach has fostered behavioral understanding and its transcoding into computational language. Undoubtedly, the investigation of brain connections and neurotransmitters in creative acts constitutes a pivotal area of inquiry to comprehend the intricacies of the human process. This will be a support for the assisted creative process, which involves the use of a learning machine to assist in the creation of art, design, and other human creative endeavors. Questions of art play a crucial role in the empathic process of exchange between the computer and the human universe. The driving force behind this connection is the assimilation of data.

The significance of artificial systems in both technological advancement and the human–computer interface cannot be overstated. These factors have the potential to foster heightened empathy. Empathic technologies are defined as technological devices that investigate biosignals to comprehend the biological mechanisms in humans, thereby acquiring information about us. When humans are exposed to biometric processes and other types

of biological signal capture, such as digital watches and smartphones, they respond with information sent by the biological signals in their brains to digital devices. These technologies are referred to as neurotechnologies. Farahany (2023), a bioethicist, lawyer, philosopher, and Iranian–American researcher, proposes that the same neurotechnologies capable of aiding neurological health, treating diseases, managing compulsions, and enhancing mental states will be sold to large technology corporations, depriving us of our mental freedom. According to her, this phenomenon is already occurring, and major technology companies are leveraging user data to influence behaviors and patterns of consumption. She asserts that in the 21st century, the preservation of cognitive freedom is paramount. The fields of machine learning and AI are undergoing rapid advancements, and existing legislation and international treaties are beginning to grant individuals even rudimentary sovereignty over their brains. The phobic reaction to AI is indicative of a narcissistic attitude. Attempts to impose control over it, whether initiated by corporations or state entities, are destined to fail. The creation and art that define human existence will only be possible if they are mediated by AI.

## *2.2 Creation with artificial intelligence*

In the domain of arts, visual and music, the utilization of AI for the purpose of artistic thought and creation first emerged in the late 1950s. Since that time, it has become a recurring theme in both the realm of art and the theoretical study of art. The initial application of AI in the arts was experimental in nature, prompting inquiries into the feasibility of comparing human and artificial thought processes. To illustrate this point, consider the use of Mondrian as a database feed with original paint. The objective of this study is to generate a pseudo-Mondrian from the painter's works. The works produced met with the public's favor. This phenomenon gave rise to numerous inquiries concerning the role of art and the artist as a creative process, a social context, and other related matters. In the book entitled *Artificial aesthetics: A critical guide to AI, media and design* (Manovich & Arielli, 2021), the authors present a diverse array of examples that illustrate the utilization of AI in creative endeavors. The text offers a comprehensive exploration of the historical and contemporary applications

of algorithms and AI, while concurrently challenging the prevailing anthropocentric paradigm concerning creativity and the arts. Consequently, the utilization of AI has been adopted by artists and art critics.

In the contemporary era, marked by the advancement of algorithms and AI technologies, these discourses have acquired a more prominent dimension. The philosopher of information, Luciano Floridi, in his book *The fourth revolution: How the infosphere is reshaping human reality* (Floridi, 2014), proposes a response to our current narcissistic crisis. This phenomenon can be attributed to the advancements in AI, its integration within major technological enterprises, and the apprehension among artists regarding their potential replacement by AI. Prior to this paradigm shift, the concepts of art and creation were considered exclusively human attributes. The awareness of potential replacement in circumstances that define our distinctiveness can influence our self-perception. This implicit defense of our exceptional place in the universe, which still existed, will collapse. We were confident that no other creature on Earth could surpass us in intelligence. The infosphere is defined as an artificial informational agent that processes information on a large scale. While such agents have not yet reached the same level of intelligence as humans, they are rapidly approaching this benchmark. Advances in imaging technology have led to a situation in which our bodies, or bio-organs, are increasingly transparent. This phenomenon is evident in a variety of imaging technologies, including video surveillance, CT scans, MRIs, ultrasounds, and neuroimaging. The preponderance of contemporary medical technologies has had a profound impact on the human body, as evidenced by the significant alterations it has undergone (Floridi, 2014).

This phenomenon can be understood as a form of narcissistic terror, stemming from the realization that we are no longer the sole beings endowed with intelligence. The field of AI has reached a point where it has surpassed human capabilities. In the context of AI being tasked with the creation of objects that exhibit aesthetic properties akin to art and other human creations characterized by creativity, a pertinent question emerges: How should we, as a society, respond to these emerging realities? In Ancient Greece, Socrates established the human being as separate from nature and, consequently, initiated humanism based on language and intelligence. The concept of humanism, founded on

"anthropological difference," was originally theorized by Socrates, who is widely regarded as the originator of the concept of man. Socrates' radical distancing from the natural world is widely considered to be the foundation for the development of humanism (Simondon, 2008). The man commences a systematic, introspective examination of his own being. If intelligence is considered a tool for the foundation of humankind, it can also be argued that it serves to put an end to anthropocentrism. Consequently, the basis for human identity, as established by linguistic differences, also signifies the dissolution of certain fundamental characteristics and pillars of the human condition. Furthermore, artistic and creative endeavors stemming from human intelligence and emotion also demonstrate a similar tendency to succumb in this process: the conclusion of the artist in the 21st century. The notion of creation in 21st-century art is a subject that merits close examination. The process of creation, when considered as an individual and subjective phenomenon, can be understood as a more visceral and physiological occurrence than a metaphysical one. As posited by Onians (2007), "subjectivity" is a more authentic phenomenon than previously theorized, being shaped less by ideologies and discourses, and more by cerebral and visceral experiences.

The following proposal will present a theoretical framework for understanding the neurobiological underpinnings of human creativity, with a focus on the role of specific neurotransmitters and cerebral regions in facilitating creative processes. Additionally, the implications of psychological faces and subjective experiences in creativity will be explored. We will propose an analysis of empirical processes because we are artists, and our proposition emerges in the practices and creative process of our students and our own creative process as artists. The study analyzes the artistic creation processes of other artists in everyday practices that could effectively impact the creative brain, and how these areas and neurotransmitters are active in art.

## *2.3 Brain steps for the creative process*

First, it is necessary to ascertain the brain regions and neurotransmitters implicated to segment the creative process into discrete phases. A synapse has the capacity to transmit signals

to multiple outputs. For instance, the primary visual cortex is responsible for processing a portion of the sensory information and relaying it to other regions of the brain: the motor cortex, the limbic system, and memory. The hippocampal apparatus is critical for the formation of explicit episodic and semantic memories, and it is believed to be associated with dreams. The neocortex is a region that stores the content of explicit memories. The amygdala plays a pivotal role in the formation of memories associated with emotions such as pleasure and fear. The basal ganglia have been demonstrated to be involved in the formation of implicit memory, which encompasses motor skills. The cerebellar region has also been implicated in implicit memory and motor learning. The prefrontal cortex is critical for short-term working memory. The phenomenon of brain stimulation is understood to occur through a specific chemical neuronal process known as the excitatory–inhibitory process. In the event that a given process is in an excitatory state, there is a reuptake of a specific kind of neurotransmitter. For example, if one situation is conducive to a depressive emotional state, the serotonin level is low, and it is captured, lowering the cortical level. This mechanism is present in every brain. When serotonin levels are high, the neuron activates the serotonin reuptake inhibitory function, thereby inducing an antidepressant state. The act of creation and the production of art invariably entail a process of recollection and remembrance. Consequently, during the creative process, acetylcholine, the neurotransmitter responsible for memory, will be inhibited. Therefore, under the fundamental principles of the creative process, the presence of these two stages is an inevitable component of any artistic endeavor. As previously mentioned, specific regions of the brain have been identified as being responsible for various cognitive functions, including memory, emotion, and motor skills. These regions have also been linked to the inhibitory and excitatory processes of neurotransmitters.

The following discussion will delineate several of the brain's stages, correlating them with the creative process. It is imperative to acknowledge that the brain does not function in a hierarchical manner; rather, it operates in a more parallel fashion. During the creative process, different stimuli and regions may be activated, underscoring the complexity of cognitive processes. Therefore, the neurotransmitter oxytocin has the capacity to elicit a narcissistic response to artistic creations. The

neurotransmitter implicated in this process has been shown to prompt artists to perceive their own work as art from an early stage (Siegel & Sapru, 2019). Semir Zeki, a neuroscientist, proposes a theory on the relationship between aesthetic pleasure, maternal love, romantic love, and the phenomenon of suspended critical judgment. According to Zeki, this suspension of judgment occurs as a result of the inhibition of the frontal and prefrontal cortex. According to the aforementioned perspective, the creative process is initiated by the artist's innate passion for their artistic creation. The term "work of art" is employed to denote an object that authentically exhibits characteristics that can be appraised as such. Second, art and creation are activities that people find enjoyable and that provide gratification and pleasure. The experience of creation evokes parallels with other pleasurable experiences. Consequently, it aligns with the principles of pleasure. In a progressive process that becomes increasingly intense for this reason, it is difficult to relinquish this pleasure. The nature of the experience, whether positive or negative, is inconsequential. Furthermore, it is imperative to encourage those who create to return to the material. Pleasure serves as a catalyst for this process. In this sense, dopamine has been linked to various experiences of pleasure, including the consumption of alcohol, drugs, sexual activities, and, notably, the aesthetic experience. This phenomenon is of interest to both the producer of the art and its appreciator.

The subjective pleasure experienced during the process of artistic creation has been demonstrated to facilitate the release of dopamine, a neurotransmitter associated with reward-related behavior. Just as the human experience of love depends on the presence of novelty to maintain dopamine levels in the brain, the human experience of art also requires new experiences to be sustained. Recent findings have revealed that dopamine's role extends beyond the realm of pleasure acquisition. This phenomenon engenders a more profound and pervasive sensation of pleasure. The objective of this study is to explore the subjective definition of pleasure. Consequently, dopamine emerges as a pivotal neurotransmitter in the regulation and prediction of behaviors. This phenomenon is observable across a wide array of human activities, including the creation of art, literature, and music; the pursuit of success; the exploration of new realms and the discovery of new laws of nature; contemplation on profound questions such as the existence of God; and the experience of romantic love. The

relationship between drug use and experiences of disappointment is a complex one. Dopamine has been shown to imply a cycle of frustration due to the necessary novelty; thus, adaptation to it is rapid. As demonstrated in the research by Lieberman and Long (2018), there has been an increase in the use of prohibited substances, including alcohol and drugs.

Consequently, the dopaminergic nature of creative activity is rooted in the inherent pleasure derived from the act of creation. Therefore, as is the case with other addictive behaviors, the experience of pleasure reemerges. The distinction in the context of creativity is that dopamine levels remained stable during these processes. This distinction between creativity and other addictive behaviors is pivotal in understanding the positive impact of creativity on the selective inhibitory cycle of dopamine reuptake. This biological dopaminergic drive then underwent a cycle in which our satisfaction became a static state, and the initial pleasure ceased to exist. The artistic creation process is cyclical, characterized by a cycle of narcissistic love for the created work, followed by a subsequent frustration that prompts a return to the artistic practice to reestablish the initial experience of pleasure and creation. The refinement of a work of art is achieved by identifying the optimal solution that will elicit a sense of pleasure in oneself and in the observer. Another important factor in art is adrenaline. The auditory cortex serves as a catalyst for adrenergic channels within the brain, thereby initiating creative processes. The process of creation, in general, and the artistic endeavor, in particular, are enhanced by a certain aggressive process.

A recent study developed by neuroscientists suggests that creativity may be enhanced in cases of frontotemporal dementia and Alzheimer's disease. From this standpoint, the notion of non-specialization in specific areas of the brain for particular tasks is contemplated (Friedberg et al., 2023). This scientific research attests to the need for the visual arts in therapies to alleviate symptoms of diseases. This prompts a pivotal inquiry: what factors underpin the observed enhancement of creativity in the context of dementia processes? If dementia is a disease that increases creative capacity in the visual arts, it may support our thesis that the creative process begins at a narcissistic stage. The initial stage in the creative process is often characterized by imitation and the suspension of critical judgment. Therefore, the existence of a particular class of neurons in the frontal and

prefrontal regions of the brain, known as mirror neurons, has been demonstrated to support certain traditional theories of art, such as mimesis. However, the mimetic process does not align with the contemporary art and creativity paradigm. In summary, we put forth a proposal for the development of an artificial creation system. The development of specialized processors and software that can substitute for the human creative cerebral mechanism is imperative. Accordingly, it can be posited that cerebral regions can be conceptualized as a distinct hardware configuration for a creative processor, with neurotransmitters functioning as the signal emitted by these regions. Regions can be considered analogous to hardware, while neurotransmitters can be regarded as analogous to software.

The subjectivity inherent in this medium can be likened to a distinct form of AI. Rather than exhausting the potential of knowledge regarding the bodily sensory reactions originating from the brain, we propose an illustration of how diverse fields must engage to generate a creative algorithm. It has become increasingly evident that interactions with the digital realm elicit behavioral modifications. The present moment is characterized by the pervasive use of social networks and search algorithms, as well as the appropriation of data from individuals by communication and information technologies. Consequently, the intervention of prominent corporations in our behaviors has become a prevailing reality. In this study, we propose a methodology for the implementation of this process through two distinct pathways. In essence, this involves the augmentation of creative capacity. In another sense, the search is underway for a truly creative entity—one that transcends the human condition and possesses the potential to augment our creative capacity.

## *2.4 Visual interfaces: A historical perspective*

Technological devices require a solution for formal and symbolic understanding to become comprehensible and understood in the context of digital culture. The image interface is the medium through which exchanges are established between the human and the artificial domains. In *A vision of the brain*, Zeki (1993) proposes that the faculty of vision is an evolutionary mechanism that facilitates the acquisition of knowledge about the world.

He proposes that the act of perceiving is a mechanism through which individuals acquire knowledge about the world. The advent of technologies such as AI, which are capable of acquiring data about individuals through biometric analysis of their eyes, marks a pivotal shift in the relationship between humans and technology. It is evident that our society is undergoing a process of increasing decoding. It is important to note, however, that certain fields within computational science have emerged as leaders in the development of these technologies. This field of computer science research has witnessed significant advancements, including the development of virtual representations and simulations of real phenomena. The real world has been expanded by computational simulations, which have radically altered its conception and perception. To investigate and create simulations of the real world using a computational approach, neuroscience is an essential field..

The field of visual computation was established in 1980 as a subdiscipline of computer science that investigates the formation of images in the primary visual cortex. These approaches have precipitated a paradigm shift in the domain of graphic interfaces in computational vision. Consequently, the advent of new technological devices is conceivable. A computational scanning of the brain has emerged as a novel paradigm for analytical processes that occur in the brain when it is exposed to art objects or aesthetic experiences. The neuroimaging of the brain has enabled the observation of neuronal activity, cerebral processes, and cortical areas that are activated in response to aesthetic sensations. The concept of neuroaesthetics was initially introduced by Zeki (1993), and it has since garnered increasing attention from researchers worldwide. The advent of new subbranches of the aesthetic process and the brain has rendered them significant objects of investigation. Neuromarketing, neuroarthistory, and neurobiological drives are perceived as fundamental to integrating humans and artificial beings. Artificial intelligence is a new paradigm for establishing a brain–interface connection. This phenomenon is exemplified by the advent of smartphones, which have profoundly impacted human behavior and lifestyle patterns in contemporary society. This phenomenon can be likened to a shift in the configuration of our brain's neural networks.

The advent of smart technologies has precipitated a period of profound transformation in human beings, exerting a significant

influence on our social relationships, affective sensibilities, and sexual partnerships. At least, the development of a more powerful computers is a new paradigm, and the human species has become interchangeable with digital technologies. In the future, artificial devices may potentially facilitate the development of frontal and prefrontal cortices, regions associated with self-consciousness and individual identity. Consequently, the development of AI will reach a point where it exhibits self-consciousness, a concept referred to as the "singularity" within the domain of AI. Throughout the history of art, creation and experimentation by artists have played a significant role in the advancement of science and technology. The creative act, in its initial phase, is characterized by the concepts of invention and reinvention. Photography is a technique that has evolved from the primordial projections in caves to the advent of photography and photochemical cinema. Consequently, the fundamental principles of visual arts have provided a conducive environment for the development of novel concepts and methodologies in the technical realm. The utilization of creation as a medium for reflecting on the technical world has been a persistent practice among artists throughout history. These revolutions have consistently sought to broaden the creative domain. Nevertheless, experiments have historically functioned as a primary catalyst for advancements in scientific knowledge.

Art, as a primarily technical and technological phenomenon, has evolved and driven the emergence of ways of representing, constructing, and perceiving the world. Since the advent of perspective, artists have altered their perception and treatment of images, shifting from fixed images on church walls to paintings that allow the image to move. The advent of the 20th century marked the emergence of photochemical photography, kinetic works, computer art, electronic art, and, most recently, digital art, along with the integration of algorithms in artistic creation with the aid of AI. A wide array of scientific and cultural domains has exerted a pioneering influence on this process, encompassing disciplines such as art and neuroscience research. In their visual experiments, artists intuitively investigated the eye as an apparatus, thereby discovering how the visual cortex functions and thus advancing science. This paradigm shift was driven by several fields, including the transition from writing code to create programming codes to behavioral psychology to contemplate AI

(Boden, 2007). At that historical juncture, the domain of human–computer interface research was undergoing significant development, with a particular focus on image-based interfaces. This project, initiated in the 1980s, continues to the present day with the use of AI. The field of research concerning the functioning of the human brain and the manner in which we comprehend the visual world has undergone rapid development. The phenomenon of digital images, with its capacity to captivate and entice, has emerged as a recurrent subject in these inquiries. This phenomenon can be attributed to the pervasive presence of screens in our daily lives and the manner in which we engage with and process digital images (Marr, 1982).

In the 21st century, significant advancements in AI have led to a resurgence of interest in the human brain, prompting extensive research in this field. The advent of neuroscience technologies has precipitated this advancement, with biosignals being the focus of considerable research. The concept of "moistware" was developed by the artist Roy Ascott, as outlined in his *Moist media manifesto* (Ascott, 2000). In his proposal, the computational machinery must become more organic to facilitate the conclusion of fusion with our bodies. Presently, this objective remains a remote aspiration within the realm of technological advancement. The human specimen is more artificial, while the computers are more organic. It is possible that this solution is idealistic and will not become a reality in the foreseeable future, particularly in the context of technological solutions in today's time. However, due to the limited availability of materials and minerals, research efforts must explore alternative solutions to produce new computational technologies. The proposed methodology aims to integrate concepts of art, neuroscience, and technology to investigate creative processes mediated by AI and visual interfaces. The objective of this study is to comprehend the manner in which technological advancements can augment the potential for artistic creation. To this end, an analysis will be conducted to study the biological and computational influences on the creative process.

# 3 METHODOLOGY: FRAMEWORK FOR CREATIVE PROCESSES IN DIGITAL ART

The proposed methodological framework integrates concepts from the arts, neuroscience, and technology to explore creative processes mediated by AI and visual interfaces. The text places significant emphasis on the concept of algorithmic literacy, asserting its indispensability for comprehending and manipulating digital media. The creative cycles, drawing inspiration from Manovich's principles of modularity and Boden's creativity levels, are methodically structured into five stages: ideation, prototyping, simulation, feedback, and finalization. This structured approach, underpinned by iterative experimentation, aims to enhance artistic possibilities. The central objective of this study is to understand how technological advancements can expand the possibilities of artistic creation. To this end, the study will analyze both biological and computational influences on the creative process. As a critical competency within this digital and algorithmic framework, algorithmic literacy is emphasized. Algorithmic literacy is defined as the capacity to comprehend, interpret, and reason about algorithms and their processes, as well as to identify their applications in both open and embedded systems. The ability to create and apply algorithmic tools and methods to solve issues across a range of fields is essential. The acquisition of an understanding of the underlying logic of algorithmic processes is integral to the development of algorithmic literacy, which enables individuals to effectively manipulate computational systems rather than being passively influenced by them. An increasing number of individuals, particularly those engaged in artistic pursuits, are advocating for the implementation of practical applications and the utilization of technology that facilitates the democratization of artistic expression through digital media (Semeler et al., 2024).

In the realm of digital art, the term "creative cycles" signifies a recurrent, iterative process entailing the evolution of concepts, their exploration, refinement, and culmination within a digital environment. This notion aligns with the perspective articulated by Manovich (2001) in *The language of new media*, wherein the author posits that digital creation is inherently modular and open to manipulation across multiple stages, facilitating iterations and real-time adjustments. According to Manovich, modularity

constitutes a foundational principle of digital media, comprising five distinct aspects: (1) numerical representation, (2) modularity, (3) automation, (4) variability, and (5) transcoding. The selection of the second principle is derived from its pertinence in the context of contemporary technological advancements. This principle is predicated on the fractal structure of new media, thereby enabling the deconstruction and reconstruction of artistic elements. Consequently, artists are empowered to experiment with different configurations and forms. This process mirrors the principles of iterative design, wherein the creative process is constantly refined through successive experimentation, embodying the concept of creative loops. Boden's (2010) seminal work, *Creativity and art: Three roads to surprise*, builds upon this perspective by exploring how computational processes not only support but also expand creative cycles. Boden proposes a taxonomy of three distinct levels of digital creativity: combinational, exploratory, and transformational. Boden emphasizes that creativity is influenced by cognitive mechanisms that can be nurtured through the acquisition of diverse knowledge, experimentation, and systematic practice within specific artistic styles.

Furthermore, she posits that cultural attitudes have the potential to impede creativity, particularly when they result in the suppression of novel and surprising ideas, thereby hindering innovation. Therefore, it is imperative to comprehend the cognitive processes underlying creative thinking to cultivate innovation in both artistic and technological domains. The fundamental understanding of algorithmic literacy can be defined as the ability to comprehend and generate sequences of instructions in a computer language that are executed to achieve a specific programming objective. The process entails the formulation of logical propositions—that is, true or false statements—through the utilization of conditions, recursion, looping, and various data structures that a computer is capable of processing. This understanding enables creators of algorithms to become proficient in any programming language, as the principles of algorithmic logic are universal across computational systems. Proficiency in algorithmic thinking empowers artists and technologists to leverage AI and visual interfaces in innovative ways, enhancing the creative process through structured experimentation and iterative refinement. Consequently, the symbiotic relationship between digital creativity and algorithmic literacy serves as the foundational

element for broadening artistic horizons in the digital era, thereby facilitating the emergence of novel forms of artistic expression and interactive design.

The stages of the creative cycle in digital art methodology was presented, which divides the creative process in digital art into five main stages, based on these theoretical underpinnings. Initial conception (ideation): At this nascent stage, the stimulation of neural activation is initiated by neurotransmitters such as dopamine and oxytocin, which are associated with inspiration and the conceptual formulation of the artistic work. This stage is characterized by the generation of preliminary concepts and the initial visualization of ideas. Visual experimentation (prototyping): The conceptual ideas are materialized within visual interfaces, thereby enabling the artist to manipulate and explore aesthetic elements in a digital environment. This phase is characterized by experimentation with various forms, colors, structures, and interactive components, which are facilitated by digital tools. The following is a discussion of technological interaction (computer simulation): The application of advanced algorithms and computational techniques is instrumental in the optimization of artistic representation. This encompasses the use of computer graphics and AI to simulate visual and behavioral effects, thereby enhancing the realism and interactive potential of the digital artwork. Creative feedback (aesthetic feedback): In this stage, the digital artwork undergoes a critical evaluation in both aesthetic and conceptual terms. Sensory and perceptual responses are analyzed, prompting adjustments and refinements to the visual representation and interactive elements of the piece. This phase is indicative of Manovich's concept of iterative design, wherein feedback loops drive continuous enhancement. Conclusion (final product): The creative process culminates with the integration of technological and artistic elements into a fully realized digital artifact. This final product represents the convergence of conceptual design, technological interaction, and aesthetic refinement, embodying the digital creative cycle's iterative and modular nature.

In this study, we propose a model that integrates aesthetics, AI, and the neuroscience of art to develop a prototype of technologies for creative applications. This approach has the potential to deepen our understanding of digital creativity and to provide artists with a structured approach to the creation of new works in a

computer-mediated environment. It achieves this by integrating neuroscience, technology, and artistic practice. The convergence of biological inspiration and digital manipulation has given rise to novel approaches in the realm of digital art, thereby extending the boundaries of what can be exhibited and experienced in digital environments.

## 4 RESULTS AND DISCUSSIONS

In the domain of modern art, the advent of the algorithmic revolution has precipitated a paradigm shift, profoundly altering prevailing notions concerning the nature of artistic creation and the role of the artist in the 21st century. The advent of computers and digital interfaces has led to a paradigm shift in the nature of the creative process, which is now performed not solely by humans but also by machines. Additionally, it encompasses AIs capable of artistic creation, thereby challenging the prevailing notion that human creativity is the exclusive domain of humans. In this study, we propose a theoretical framework that explores the notion that art constitutes a distinctively human form of expression that has come to the fore in the age of algorithms, capable of replicating both cognitive and aesthetic functions. This shift in perspective entails a reconfiguration of the relationship between art and technology, giving rise to a novel paradigm of "computational mannerism." Within this paradigm, aesthetic production is influenced by algorithmic processes that emulate creative behaviors. The historical analysis presented in the document demonstrates that art has utilized scientific advancements to broaden its expressive capabilities since the advent of the 20th century. This convergence was initiated by the incorporation of psychoanalytic theories into Dadaism and Surrealism. The advent of computer systems modeled on the human brain has precipitated the proliferation of AI as a creative instrument.

This technological advancement has facilitated the materialization of abstract concepts within visual interfaces, thereby effecting a transformation in the perception of the creative process and integrating the algorithm as a collaborative agent in artistic creation. Furthermore, an understanding of the biological mechanisms underlying creativity is imperative, and neuroscience is instrumental in elucidating these mechanisms.

Neuroimaging studies have demonstrated the activation of specific brain areas, such as the limbic system and the visual cortex, during the process of creative thinking. It has been posited that the relationship between neurotransmitters such as oxytocin and dopamine plays a pivotal role in artistic motivation and the experience of aesthetic pleasure. The integration of this biological mapping into the development of algorithmic interfaces has enabled machines to imitate artistic experiences in a manner that is increasingly autonomous and intricate. Furthermore, emphasis has been placed on the notion that the concept of visual computing signifies a pivotal moment in the evolution of digital art creation. This technique, developed in the 1980s, has expanded the potential for computer art by simulating natural phenomena in virtual environments. Advancements in particle simulation, advanced graphic interfaces, and facial recognition technologies have enabled the representation of reality in a manner that is unparalleled, thereby solidifying the role of AI as a creative agent. Consequently, the theoretical analysis' findings suggest that the algorithmic revolution has not only revolutionized the production of art but has also posed philosophical questions regarding the artist's role and the essence of creativity. By incorporating biological and cognitive mechanisms, the algorithm emerges as a prominent figure in contemporary creation, signifying a future in which the distinction between human and machine becomes progressively indistinct in the artistic domain.

## 5 CONCLUSION

In conclusion, although the integration of art, AI, and neuroscience has advanced significantly, it is unclear to what extent AI-mediated creative processes can be considered authentic. The repercussions of this phenomenon on the conceptualization of authorship, aesthetic value, and creative consciousness remain to be elucidated. The dearth of consensual criteria for evaluating artificial creativity gives rise to epistemological and ethical questions concerning the entity responsible for the creation. Consequently, the issue of what factors contribute to the designation of a creation as genuinely creative arises. The role of emotion and intentionality in this process is a critical question that must be addressed. The integration of art, AI, and neuroscience has led

to a redefinition of the concept of creativity, thereby disrupting the long-standing anthropocentric paradigm that has sustained the idea for centuries, asserting that artistic creation is an exclusively human phenomenon. Digital technologies, propelled by advancements in neuroimaging and machine learning algorithms, have demonstrated the capacity to replicate and augment creative processes through artificial systems.

Through the implementation of algorithmic simulations and the integration of enhanced visual interfaces, it becomes evident that the materialization of aesthetic concepts is not merely an outcome of the process but rather a crucial element in itself. Moreover, the emulation of cognitive processes associated with the creative act is not merely a byproduct but an intentional component of the design. Recent neuroscientific studies have indicated that the activation of specific regions of the brain, such as the prefrontal cortex and limbic system, in conjunction with neurotransmitters such as dopamine and oxytocin, plays a critical role in the experience of beauty and the motivation for creative endeavors. This understanding facilitates the development of AI systems that replicate these mechanisms, thereby enabling artistic creation that transcends human intentionality. From this vantage point, the advent of AI-mediated artistic creation has given rise to a novel domain of theoretical and philosophical inquiry, one that interrogates the boundaries between authorship, originality, and creative awareness. The capacity of machines to engender works that evoke intricate emotional and aesthetic responses necessitates a reevaluation of the conventional notions of art and the artist. This reevaluation suggests a transition from a human protagonist to a hybrid cognition shared with artificial devices.

This phenomenon not only alters the process of creation but also impacts the manner in which we interpret and value artistic production. Consequently, the evolution of AI technologies applied to art represents more than a mere technical advancement; it is an invitation to reinterpret the aesthetic and ontological foundations of creation. The boundaries between human and machine become increasingly indistinct, and the concept of creativity expands, paving the way for an era in which art and technology coexist and influence each other and redefining the very meaning of creation. In contemplating creativity within a paradigm characterized by the pervasive apprehension of being

superseded by AI, it becomes evident that there is an imperative and urgent need for augmented investment and research in this domain. Innovation, defined as the introduction of new ideas or methods, permeates all fields of knowledge, and the basis for this is creativity. At present, we are observing a "simulation of human creativity." However, this does not imply that AI will not eventually exceed human capabilities in this domain.

In conclusion, prospective endeavors in the sphere of AI system development include the conceptualization of methodologies that facilitate the emulation of affective states during the creative process. These methodologies are predicated on the utilization of dopamine and oxytocin models. The objective of this study is to examine the design of hybrid creative platforms that facilitate collaborative authorship between humans and AI agents. The following study will explore the legal frameworks that have been established for the management of intellectual property in the domain of AI-generated art. This study presents the findings of longitudinal studies on the evolving public perception of AI art over time.

## *Conflict of interest*

The authors of this article declare that they have no conflict of interest.

## *Contribution statement*

**Conceptualization:** Alberto Marinho Ribas Semeler, Alexandre Ribas Semeler.
**Data Curation:** Alberto Marinho Ribas Semeler, Alexandre Ribas Semeler.
**Formal Analysis:** Alberto Marinho Ribas Semeler, Alexandre Ribas Semeler.
**Methodology:** Alberto Marinho Ribas Semeler, Alexandre Ribas Semeler.
**Writing – Alberto Marinho Ribas Semeler, Alexandre Ribas Semeler.**
**Writing – Review and Editing:** Alberto Marinho Ribas Semeler, Alexandre Ribas Semeler.

### *Statement of data consent*

The data generated during the development of this study have been included in the manuscript.

## REFERENCES


Boden, M. (2007). *Creativity: How does it work?* University of Sussex.

Boden, M. A. (2010). *Creativity and art: Three roads to surprise*. Oxford University Press.

Farahany, N. A. (2023). *The battle for your brain: Defending your right to think freely in the age of neurotechnology*. St. Martin's Press.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality* (Kindle ed.).

Friedberg, A., Pasquini, L., Diggs, R., Diggs, R., Glaubitz, E. A., Lopez, L., Illán-Gala, I., Iaccarino, L., La Joie, R., Mundada, N., Knudtson, M., Neylan, K., Brown, J., Allen, I. E., Rankin, K. P., Bonham, L.W., Yokoyama, J. S., Ramos, E. M., Geschwind, D. H. ... Miller B. L. (2023). Prevalence, timing, and network localization of emergent visual creativity in frontotemporal dementia. *JAMA Neurology*, *80*(4), 377–387. *https://doi.org/10.1001/jamaneurol.2023.0001*

Kristeva, J. (1982). *Powers of horror: An essay on abjection*. Columbia University Press.

LeDoux, J. (2015). *Anxious: Using the brain to understand and treat fear and anxiety* (Kindle ed.). Penguin.

Lieberman, D. Z., & Long, M. E. (2018). *The molecule of more: How a single chemical in your brain drives love, sex, and creativity—and will determine the fate of humanity* (Kindle ed.).

Manovich, L. (2001). *The language of new media*. MIT Press.

Manovich, L., & Arielli, E. (2021). *Artificial aesthetics: A critical guide to AI, media and design*. Self-published.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Onians, J. (2007). *Neuroarthistory: From Aristotle and Pliny to Baxandall and Zeki*. Yale University Press.
Paul, C. (2015). *Digital art* (3rd ed.). Thames & Hudson.
Semeler, A., Pinto, A., Koltay, T., Dias, T., Oliveira, A., González, J., & Rozados, H. B. F. (2024). Algorithmic literacy: Generative artificial intelligence technologies for data librarians. *EAI Endorsed Transactions on Scalable Information Systems, 11*(2). *https://doi.org/10.4108/eetsis.4067*
Siegel, A., & Sapru, H. N. (2019). *Essential neuroscience*. Wolters Kluwer.
Simondon, G. (2008). *Dos lecciones sobre el hombre y el animal*. Cebra.
Vartanian, O., Bristol, S., Adans, C., & Kaufman, J. C. (2013). *Neuroscience of creativity*. MIT Press.
Zeki, S. (1999). *Inner vision: An exploration of art and the brain*. Cambridge University Press.

CHAPTER 2

# OBSTETRIC DECISION-SUPPORT SYSTEM: AN INFORMATIONAL MODEL FOR MATERNAL AUTONOMY TOWARDS THE AGENDA 2030 HEALTH GOALS

**Paulianne Fontoura Guilherme de Souza**
*Department of Information Science, Federal University of Santa Catarina (UFSC), Brazil.*
ORCID: *https://orcid.org/0009-0009-5638-9274*

**Gustavo Geraldo de Sá Teles Junior**
*Postgraduate Program in Philosophy, Federal University of Santa Catarina (UFSC), Brazil.*
ORCID: *https://orcid.org/0000-0002-1459-2014*

**Douglas Dyllon Jeronimo de Macedo**
*Department of Information Science, Federal University of Santa Catarina (UFSC), Brazil.*
**Email:** *douglas.macedo@ufsc.br*
ORCID: *https://orcid.org/0000-0002-3237-4168*

## ABSTRACT

This study proposed an obstetric decision-support system with the objective of strengthening maternal autonomy and advancing the health objectives of Agenda 2030, with a particular focus on Sustainable Development Goal 3. The research employed a

qualitative, applied, and exploratory approach, integrating digital health, evidence-based care, and epistemological perspectives to formulate a structured environment tailored to pregnant women. The objective of this study was to address critical informational asymmetries in obstetric care through a digitally enabled, evidence-based model that empowered pregnant individuals in clinical decision-making. The methodological approach entailed a comprehensive review of the extant literature, requirement engineering, and the application of Unified Modeling Language diagrams to formalize system functionalities. The results of the study included a multilayer informational infrastructure comprising a dynamic layer for adaptive learning via algorithmic curation of scientific evidence and lived experiences, enabling personalized recommendations, and a static layer for birth planning and scenario simulation, integrating validated guidelines to reduce uncertainty. The integration of scientific evidence with user experiences facilitated the translation of World Health Organization guidelines into comprehensible and implementable information, thereby reducing informational asymmetries and enhancing the gestational decision-making process. The architecture prioritized intelligibility, traceability, and user-centric navigation, ensuring alignment with maternal profiles and preferences. The findings suggested that the formal modeling process facilitated the conversion of normative content into a computable and auditable structure, thereby promoting autonomy without compromising clinical safety. The study's findings indicated that a digitally mediated informational base, conceptualized at the nexus of health, information, and technology, served as a pivotal infrastructure for ensuring dignified, equitable, and humanized obstetric care. This contributed structurally to high-quality health systems by bridging evidence–practice gaps.

---

**KEYWORDS**: information systems, evidence-based care, health humanization, UML modeling, obstetric decision

**HOW TO CITE**: Fontoura Guilherme de Souza, P., Teles Junior, G. G. de S., & Dyllon Jeronimo de Macedo, D. (2025). Obstetric decision-support system: An informational model for maternal autonomy towards the Agenda 2030 health goals. In A. Semeler (Ed.), Artificial Intelligence and Data Science Practices in

Scientific Development, *Advanced Notes in Information Science, volume 8* (pp. 40-72). Pro-Metrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9331-4-5.53.

**COPYRIGHT**: © 2025 The author(s). This article is distributed under the terms of the CC BY-NC 4.0 license, which permits copying and redistribution of the material in any medium or format, adaptation, transformation, and building upon the material, provided that the license terms are followed.

## 1 INTRODUCTION

According to the World Health Organization (WHO), health is defined as a state of complete physical, mental, and social well-being that transcends the mere absence of disease or infirmity (WHO, 1946). This conception serves to expand the understanding to a biopsychosocial perspective, which is fundamental in the formulation of the Sustainable Development Goals (SDGs), especially SDG 3: "Ensure healthy lives and promote well-being for all at all ages" (UN, 2015, p. 23). Examples of this evolution include telemedicine and telehealth approaches (Macedo et al., 2015; Puel et al., 2014; Soares et al., 2013; Souza Inácio et al., 2014). Despite the potential of technological advancements to enhance health promotion, extend life expectancy, alleviate suffering, and promote cellular restoration, healthcare systems have historically fallen short in providing universal, safe, and high-quality care (Institute of Medicine US, 2001, pp. 2–3). Facilities equipped with structural and human resources, including health professionals, potable water, electricity, and medications, are essential for the realization of quality care. However, the isolated presence of these factors does not guarantee excellence in care (Kruk et al., 2018, p. 1197). Despite the availability of adequate tools, adverse scenarios persist, reflecting a series of failures. These failures include the increase in inappropriate, unnecessary, or out-of-context actions, which highlight significant gaps in areas such as user experience, competence, and trust in the system, as well as people's well-being.

In the domain of obstetrics, the Global Strategy for Women's, Children's, and Adolescents' Health (WHO, 2015) emphasizes the reduction of preventable mortality in addition to promoting well-being and psychological safety. This strategy is consistent

with the objectives outlined in SDG 3, which aims to ensure that "women and babies not only survive childbirth complications, should they occur, but also thrive and realize their potential for health and life" (PAHO, 2019, p. 1). This approach is intrinsically linked to human rights, as recognized in Resolution 18/2 of the Human Rights Council, which legitimizes women as active agents in decision-making regarding their sexual and reproductive health (OHCHR, 2012, p. 5). Despite the advancements in reducing maternal and neonatal mortality associated with access to facilities for high-risk cases, there has not been commensurate progress in the humanization of care, as evidenced by the proliferation of obstetric violence. This reality indicates that, during periods of heightened physical and emotional vulnerability (e.g., pregnancy, childbirth, and the postpartum period), many women encounter adverse experiences in healthcare settings, including mistreatment, disrespect, and neglect (D'Oliveira et al., 2002). The consequences of these experiences can persist not only for the affected women but also for their family units.

Moreover, the pervasive medicalization of childbirth has eroded women's capacity to autonomously manage their labor. This phenomenon is evidenced by the fact that a substantial proportion of healthy pregnant women are subjected to a range of routine clinical interventions, the efficacy of which is often questionable, and whose potential adverse effects should not be overlooked (Coulm et al., 2012; EURO-PERISTAT, 2013; Renfrew et al., 2014). The absence of a universal consensus on the definition and measurement of these issues, in conjunction with the insensitivity of interventionist methodologies to women's needs, values, and preferences, undermines informational equity and exacerbates disparities in accessing adequate care (PAHO, 2019, pp. 1, 8). It is imperative to acknowledge that all women possess an inherent right to the pinnacle of attainable well-being, encompassing dignified and respectful care, along with safeguards against violence and discrimination. Substandard practices, therefore, represent egregious violations of fundamental human rights (WHO, 2014, pp. 1–2). As indicated by the Royal College of Obstetricians and Gynecologists (2017) and Downe et al. (2015), a positive experience is contingent upon the presence of certain essential elements. These elements include the maintenance of physical normality for both the mother and the infant, as well as the prevention and treatment of risks, diseases, and death. Additionally,

the maintenance of psychosocial normality is imperative, ensuring a pregnancy characterized by self-esteem, competence, and autonomy. In this regard, the WHO has underscored the significance of imparting efficacious communication concerning physiological, biomedical, behavioral, and sociocultural concerns, along with the provision of comprehensive support, encompassing social, cultural, emotional, and psychological dimensions, to expectant mothers in a manner that respects their dignity (WHO, 2015). However, studies demonstrate that in low- and middle-income countries, adherence to evidence-based guidelines remains suboptimal, with professionals exhibiting an average compliance rate of 47% in complying with care recommendations (Kruk et al., 2018, p. 1203).

In Brazil, public policies such as the National Program for the Humanization of Childbirth (2000), the Companion Law (2005), and the "Stork Network" (Rede Cegonha, in Portuguese) (2011) were implemented in an effort to address these challenges. However, the official exclusion of the term "obstetric violence" from policy guidelines in 2019 (Leite, 2014, 2021) compromised progress in measuring and addressing these practices. In this scenario, access to qualified information is configured not only as a right but as an indispensable condition for the empowerment of pregnant women (Targino, 1991, p. 155), thereby shifting them from passive recipients to active protagonists in their healthcare (Almeida Junior, 2009). Consequently, effective information, as a foundational element, empowers the pregnant woman to assume a central role in her gestational cycle. This empowerment enables her to seek answers that can support decisions regarding actions and interventions with a minimal degree of uncertainty, fostering the potential for argumentation and inquiry. Consequently, the dissemination of accurate information has emerged as a strategic element to mitigate vulnerabilities and enhance the humanization of obstetric care. This underscores the necessity for robust informational infrastructures that facilitate knowledge sharing and the monitoring of care quality. While equity transcends the technological sphere, digital solutions have the potential to reduce informational gaps in care for pregnant women, thereby promoting the empowerment of patients, families, and communities (Guimarães & Silva, 2011, p. 3553).

In light of this necessity, it is acknowledged that the daily engagement in health-promoting practices is contingent upon

the availability of information through communication channels, with technological advancements playing a progressively pivotal role in facilitating these activities. It is therefore emphasized that the establishment of robust foundations for an advanced health scenario necessitates not only physical tools, such as equipment, medications, and materials, but also novel attitudes, competencies, and behaviors, predicated on the capacity and propensity to learn from data (Kruk et al., 2018, p. 1202). In an era marked by accelerated digital transformation, propelled by the interplay between technological advancements and sociocultural shifts, there emerges a pronounced imperative for seamless integration of the health sector with information and communication technologies (ICTs). This integration is pivotal in attaining the objectives of paramount quality and ensuring individualized safety measures. Moreover, it is essential for fulfilling the obligations of investigation, notification, and humanitarian action in the context of public health (WHO & ITU, 2012). Digital health, defined as the systematic use of ICTs to strengthen informed decisions and promote well-being, has consolidated itself as an essential tool in care delivery (WHO, 2019). This initiative has garnered recognition from the World Health Assembly (2018), underscoring its significance in promoting universal coverage and enhancing service quality. However, the implementation of these technologies faces challenges, such as subjectivity in interpretation by developers and the absence of systematized documentation, which complicates replication and monitoring (WHO, 2021).

Accordingly, this study is guided by the following central question: By what means may an informational proposal be developed that would enhance the empowerment of those involved in obstetric care through the dissemination of evidence-based information? Answers are sought for the translation of scientific knowledge into accessible language, the prioritization of the most relevant themes for building maternal knowledge, and the integration of necessary components to ensure effective and optimized access to information. The objective of this study is to contribute to obstetric care that is more humane, equitable, and centered on women's autonomy. The study will articulate structural, social, and technological dimensions to strengthen health systems in facing contemporary challenges. In light of contemporary demands for integrated solutions between health and technology, this study proposes an obstetric informational model.

This model constitutes a structured and specialized channel. The purpose of this channel is to foster the autonomy of pregnant women. It does so by providing access to information that has been rigorously validated and updated. The model is supported by the WHO international regulations and articulated on the foundations of digital health. It adopts an informational empowerment approach, prioritizing the strengthening of the user's decision-making capacity throughout pregnancy.

The objective of this study is to develop a structured obstetric information base through an autonomous digital learning system, with the aim of promoting informational empowerment and the generation of maternal knowledge. This objective is intended to strengthen the autonomy of pregnant women and reduce uncertainty related to the relevance, consistency, and adequacy of health practices. To this end, this study seeks to analyze the specific informational demands of pregnant women, their impacts on the care process, and the necessary actions to minimize difficulties faced during pregnancy. This analysis will consider the importance of accessible and evidence-based information to transform users into active protagonists (Almeida Junior, 2009; Targino, 1991). This analysis proposes the development of a digital platform to centralize and organize content in accordance with scientific guidelines. This platform is intended to complement, rather than replace, medical guidance, in alignment with the principles of digital health and the humanization of care (Kruk et al., 2018; WHO & ITU, 2012). The successful establishment of a high-quality healthcare system is contingent upon the implementation of universal rights through the meticulous execution of evidence-based practices, disseminated by means of reliable informational flows. In the field of obstetrics, achieving excellence entails more than merely providing advanced technologies; it necessitates a gestational journey founded on the principles of women's autonomy and active engagement in clear communication processes unencumbered by biases or distractions that might compromise information integrity. The proposal's scope extends beyond the mere reduction of recognized violations, aiming to transform the health field by minimizing informational failures with the support of digital technologies.

High-quality health systems should consider individuals, families, and communities as active partners whose needs and preferences should shape institutional responses (Kruk et al.,

2018). In the context of obstetrics, the asymmetrical distribution of power and information between healthcare professionals and pregnant individuals underscores the urgency of this centrality. This dynamic functions as a moral and practical mechanism for empowerment and accountability (Kruk et al., 2018). Consequently, there is an explicit demand for informational bases that not only expand sharing networks but also translate and clarify content in an accessible way, promoting the empowerment of pregnant women. As Araújo (1992) emphasizes, the dissemination of information and knowledge has the capacity to disrupt historically discriminatory power relations. This vision is consistent with the 2030 Agenda, which emphasizes obstetrics centered on the woman as an active agent of her own health. In the subsequent section, an examination of extant literature will be conducted to explore the informational bases in health and their relationship with the humanization of obstetric care.

### *1.1 Literature review*

In recent decades, there has been a growing consensus that sustainable health advancements require not only specific technological innovations but systemic transformations that integrate social, political, economic, and cultural aspects in favor of biopsychosocial well-being. While the millennium development goals (MDGs) have been acknowledged to have promoted significant advancements, such as a 45% reduction in global maternal mortality between 1990 and 2015 (Brizuela & Tunçalp, 2017), the current 2030 Agenda has been regarded as an expansion of this scope by incorporating equity as a structuring axis. This incorporation ensures not only access but also the quality and adequacy of care. Consequently, it is acknowledged that high-performance health systems must be cognizant of individual, cultural, and contextual variations, providing ethical and effective responses to the population's demands. This global movement is exemplified by initiatives such as the Commission on High Quality Health Systems in the SDG Era, which delineates four strategic fronts: governance with a focus on quality, redesign of service delivery, transformation of the workforce, and stimulation of demand for quality by users (Kruk et al., 2018). Concurrently, the Institute of Medicine proposes six fundamental dimensions for patient-centered care:

safety, effectiveness, timeliness, efficiency, equity, and respect for individual preferences (Institute of Medicine US, 2001). By underscoring the imperative for the user experience to inform all clinical decisions, this framework underscores the pressing need to reconfigure systems to prioritize technique, active listening, and the establishment of mutual trust.

This paradigm shift is also expressed in the 10 rules for redesigning care systems proposed by the same committee. These rules include recognizing the patient as the source of control, the unrestricted sharing of information, and evidence-based decision-making (Institute of Medicine US, 2001). These guidelines underscore the notion that information is not merely an adjunct to care, but rather a fundamental element that provides a framework for, substantiates, and evaluates the quality of care. In this context, obstetrics emerges as a field replete with tensions and possibilities. The WHO has outlined guidelines for maternal and neonatal care, proposing a theoretical model based on Donabedian's (2005) structure-process-outcome triad, which is distributed into eight domains that articulate technical quality and subjective experience (WHO, 2018). Consequently, the concept of excellence in health is understood to encompass two fundamental aspects: clinical competence and the communicational and relational capacity of healthcare professionals. In this scenario, the evaluation of quality must extend beyond objective outcomes, incorporating criteria such as the adequacy, integrity, and intelligibility of the information conveyed to the user. The necessity for mechanisms that translate scientific knowledge into comprehensible language is underscored by the incorporation of information as an axis of guidance and the maturation of care. This process restores the right of pregnant women to comprehend, interrogate, and determine. As Kruk et al. (2018) assert, the establishment of a quality system is predicated on the presence of informed, engaged, and respected subjects. However, the transfer of decision-making to professionals and institutions engenders an informational asymmetry that impedes free and informed choices, particularly regarding the type of birth and interventions performed (Zorzam, 2013). This context is further exacerbated by a pathologizing conception of the female body, which transforms childbirth into a risk to be controlled by often unnecessary procedures, contradicting evidence-based recommendations (Diniz, 2009; Zorzam, 2013).

The persistence of obsolete practices, such as episiotomies, inductions without clinical indication, and risky maneuvers, further compromises the quality of obstetric care and produces deleterious effects on maternal health and well-being (Diniz, 2009). The dearth of lucid and sufficient information regarding procedures undermines autonomy, as evidenced by the finding that a mere one-third of women feel adequately informed about the exams or medications administered during childbirth (Domingues et al., 2004). Moreover, the challenge of comprehending medical counsel affects over 70% of pregnant women receiving primary care (Mota et al., 2015). This predicament engenders a milieu of uncertainties and trepidations, thereby impeding women's capacity to engage proactively in decision-making processes concerning their health. It is important to acknowledge that the absence or distortion of information can be considered a form of informational violence. This type of violence undermines the dignity and rights of pregnant women, impeding their capacity to identify abusive practices (Russo & Carrara, 2015). Confronting this predicament necessitates the establishment of obstetric informational ecosystems that integrate the equitable distribution of resources with policies that promote women's active engagement in shaping digital infrastructures, thereby facilitating communicative processes that elucidate the biological, psychological, and social dimensions involved in the construction of care (Souza et al., 2011).

## 2 METHODOLOGY

This study is configured as an Application or Technological Adaptation Project, as it is directed towards the creation of innovation assets through the generation of products, processes, devices, and services based on scientific knowledge and formatted as technological systems subject to testing and evaluation (Fuck & Vilha, 2011). Its classification as an applied investigation is predicated on its objective of achieving practical applications, with the aim of resolving specific problems (Silva & Menezes, 2001, p. 20). In the context of this work, the application manifests itself in the proposition of a model intended to generate knowledge and to instrumentalize the exchange of information about obstetric care, seeking to solve demands and fill concrete informational

gaps. With respect to its objectives, this study is exploratory in nature, as it seeks to meticulously examine facts, phenomena, or new knowledge about which there is a paucity of information (Tobar & Romano Yalour, 2001). Concurrently, it is descriptive in nature, by presenting a series of information about the object of analysis, detailing the facts and phenomena of a particular reality (Triviños, 1987). In terms of methodology, the study is classified as qualitative. This is due to the fact that the research is characterized by an inductive approach to data analysis, whereby the researcher derives concepts, ideas, and understandings from patterns identified in the studies (Reneker, 1993). The qualitative approach is applicable to the entire development process of the proposed model, encompassing its definition, conception, and subsequent model evaluation stage.

The research methodology encompasses a bibliographic survey and the development of Unified Modeling Language (UML) diagrams (Hamilton & Miles, 2006). The scientific literature provided the conceptual basis on health, information, and technology, which was crucial for identifying the problem, defining objectives, and constructing the theoretical framework. To develop the model, we utilized the Digital Adaptation Kit for Antenatal Care, the Digital Implementation Investment Guide, and the recommendations on digital interventions for health system strengthening from the WHO. These resources were complemented by principles of Requirements Engineering and Design Science Research. In the context of the investigation, the initial focus was on assessing the informational needs of pregnant women, taking into account the challenges they encounter in accessing reliable and current healthcare guidance. This preliminary diagnosis was predicated on informal data collection on digital platforms, such as social networks, and in direct conversations with pregnant women, who shared authentic accounts of their difficulties in finding information. This preliminary understanding proved to be pivotal in delineating the scope of this study, which focused on understanding a scenario characterized by uncertainties and an abundance of conflicting information.

The theoretical framework of the study was developed through an extensive review of the extant literature. The survey provided the conceptual foundation for the subsequent development of the work, in addition to delineating the general and specific objectives that guided the subsequent phases. A

documentary mapping of past and current guidelines on health, information, and obstetrics was carried out, with a special focus on reports issued by the WHO in the last decade, prioritizing the most recent documents. In addition, a comprehensive exploration was conducted of prominent databases, including Latin American and Caribbean Health Sciences Literature (LILACS), Nursing Databases (BDENF), Scientific Electronic Library Online (SciELO), ScienceDirect, Cumulative Index to Nursing and Allied Health Literature (CINAHL), and the US National Library of Medicine—National Institutes of Health (PubMed). The utilization of standardized descriptors, as delineated by DeCS (Health Sciences Descriptors) and MeSH (Medical Subject Headings), such as "health," "obstetrics," "information," "technology," and "mobile applications," was employed to conduct a comprehensive and precise data prospection. This approach was designed to encompass the multidimensional complexity of the object under analysis. Complementary searches were conducted in national repositories, such as the portal of the Coordination for the Improvement of Higher Education Personnel (CAPES), and in normative documents, manuals, guidelines, and legislation pertaining to prenatal care in Brazil, originating from the Ministry of Health. Thus, the technical-legislative aspects were articulated with the practical and informational requirements of pregnant women.

The technical specification phase was guided by the guidelines of the Digital Adaptation Kit for Antenatal Care (WHO, 2021), implemented sequentially according to the digital structuring stages recommended in said document. The functional design of the proposed system was realized using the Figma tool, grounded in the precepts of Requirements Engineering and the guidelines of Design Science Research. This approach aimed to ensure integration between functional requirements and the specificities of the obstetric domain. The modeling employed UML notation to delineate both the behavioral and structural aspects of the system, through use case, activity, sequence, class, and component diagrams. These diagrams enabled the graphical representation of the static topology of the involved entities, their interrelations, attributes, and operational flows, culminating in a logical architecture congruent with the proposition of informational empowerment through access personalization and adaptive navigability.

## 3 RESULTS

In light of the multifaceted nature of contemporary obstetric demands, a systemic proposal has been formulated with the objective of developing a multifaceted instrument capable of translating evidence and experiences into operational resources. The delineated structure is predicated on stratified modeling, conceived from an interdependent logic of distributed networks, in which each element coexists in a collaborative regime, favoring both automated learning and evidence-based decision-making. This architecture was organized into two main navigation layers: the dynamic layer and the static layer. These layers are intercomplementary in guiding the gestational journey. Consequently, they establish a modular, responsive, and evolutionary system that accommodates distinct degrees of autonomy and complexity, according to the user's informational needs and gestational stage.

### *3.1 Dynamic layer: Monitoring by science–experience*

This layer, which constitutes the active core of the system, functions as an adaptive learning interface. It is fed by scientific sources and empirical records, thereby enabling the automated screening, curation, and categorization of recommendations, rights, and narratives. This dynamism is operationalized by algorithmic extraction and filtering technologies, resulting in responsive personalization of guidance according to the user's progress on the platform. Its functional flows are structured into four axes:

1. Profile segmentation: It performs user classification based on clinical–contextual variables, allowing for accurate contextualization of recommendations.

2. Evidence mining and translation: It promotes the automated collection of clinical data and guidelines, with subsequent conversion into accessible language, maintaining scientific rigor.

3. Experience curation: It enables the sharing of experiences among users by similarity, constituting an empirical support environment.

4. Guidance personalization: It aggregates data from previous modules, promoting adapted recommendations that evolve according to the pregnant woman's journey.

### *3.2 Static layer: Planning for decision-making*

The static layer facilitates long-term planning by means of the organization of birth plans and outcome simulation. Notwithstanding the existence of non-digital records, its integration with scientifically validated guidelines enhances decision-making and reinforces user autonomy. This layer is composed of:

1. Birth plan: An interactive interface allows for detailed configuration of birth preferences, with automated recommendations based on previously mined guidelines.

2. Scenario simulation: It generates representations of the potential outcomes of choices made, accompanied by risk–benefit analysis with a clinical, logistical, and emotional focus.

### *3.3 System details*

To facilitate the integration of clinical guidelines into digital public health environments, the WHO has developed a set of digital adaptation kits. These kits are designed to standardize the translation of recommendations into operational structures that have been validated for their effectiveness (Tamrat et al., 2022). In accordance with the eight proposed structuring components delineated by the WHO (2021, p. 7), this study examines two of these components in particular. The following elements are of particular relevance in this context: operational processes and workflows and the main dataset. The inherent complexity of system development, irrespective of whether the development is for a mobile application or a corporate-scale initiative, poses challenges regarding the definition of components, their functions, and relationships. In this sense, modeling can be regarded as a

method of abstraction, with the objective of elucidating fundamental principles and facilitating technical comprehension that is amenable to critical scrutiny (Hamilton & Miles, 2006, p. 23). To ensure precise formal representation, the UML was adopted, given its capacity for modular adaptation according to the project context. In this investigation, two dimensions were prioritized: the behavioral, focusing on inter-object flows in the context of operational processes and workflows, and the structural, dedicated to the topology of intra-object data in the main dataset.

### 3.3.1 Operational processes and workflows

Operational processes encompass the systemic logic of the application's functioning in its temporal, distributed, and interactional dimensions. It entails the formalization of connections between actors, functional modules, and computational mechanisms. When articulated under defined conditions, these connections organize the solution's behavior throughout usage cycles. The following diagrams illustrate the articulation in question, meticulously unfolding the system's behavior into its constituent sublayers. This approach facilitates a comprehensive and nuanced visualization of systemic interactions (OMG, 2011).

#### *3.3.1.1 UML use case diagram*

The use case diagram elucidates the core functionalities attributed to the user and the system, delineating the contact points between them through interaction flows. The model employs dashed lines to symbolize the bidirectional transit of data and events, conceptualizing the user as an agent of actions such as navigation, data insertion, and birth plan construction. In response, the system executes tasks of data verification, mining, and validation, as well as the organization and storage of plans and experiences. The capacity for scenario simulation and access to reports is also integrated into the represented functional matrix (Figure 1).

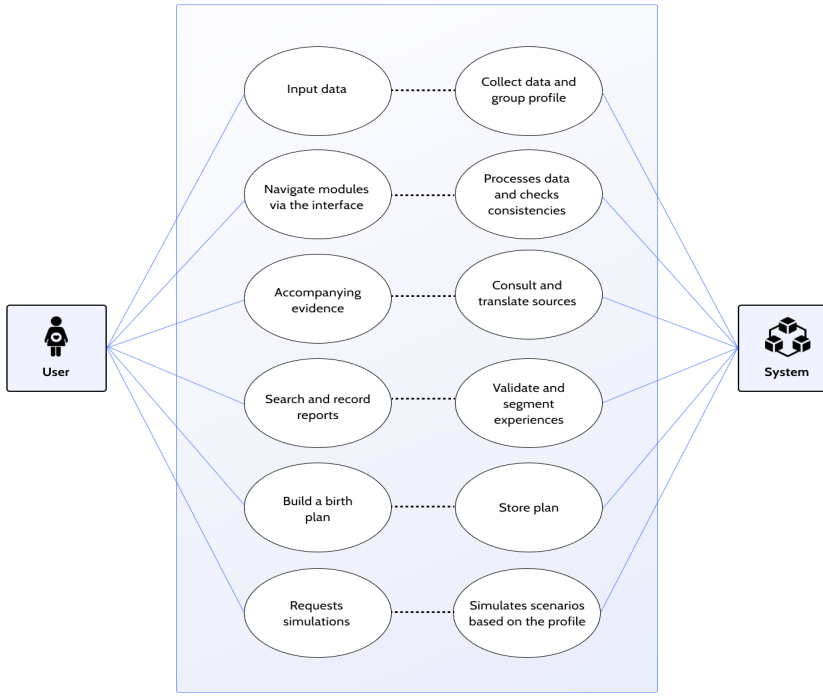


**Figure 1.** UML use case diagram. **Source**. Author.

### *3.3.1.2 UML activity diagram*

The diagram organizes the user's and system's activity modules into horizontal pools, subdivided into specific lanes for operations such as data verification, plan construction, evidence classification, and scenario analysis. The sequence of actions commences with user authentication and unfolds according to the user's choices. In this process, data are encrypted and directed to analytical engines that apply rules for generating personalized recommendations. The model also incorporates feedback cycles and continuous reconfiguration of guidance, ensuring dynamic adherence to inserted data (Figure 2).

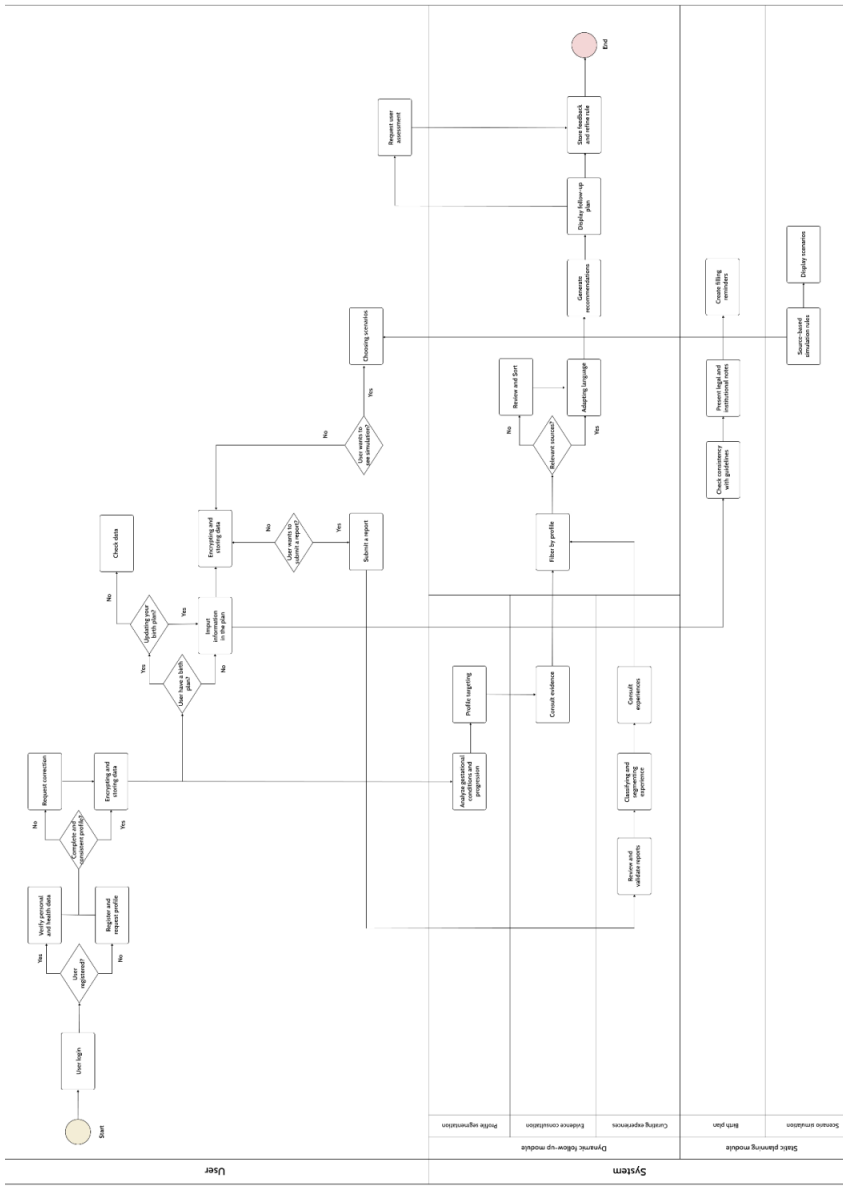


**Figure 2.** UML activity diagram. **Source**. Author.

### 3.3.1.3 *UML sequence diagram*

The sequence diagram, which is organized vertically into "lifelines," portrays the interaction among five main entities: user, interface, database, application, and rules engine. The user inputs data, which traverses the technical circuit from the interface to the database, undergoing encryption, grouping, and validation. The application processes the profiles, and the rules engine acts in the semantic verification of sources and in the classification of reports. This results in specific recommendations and iterative simulations that feed back into the system (Figure 3).

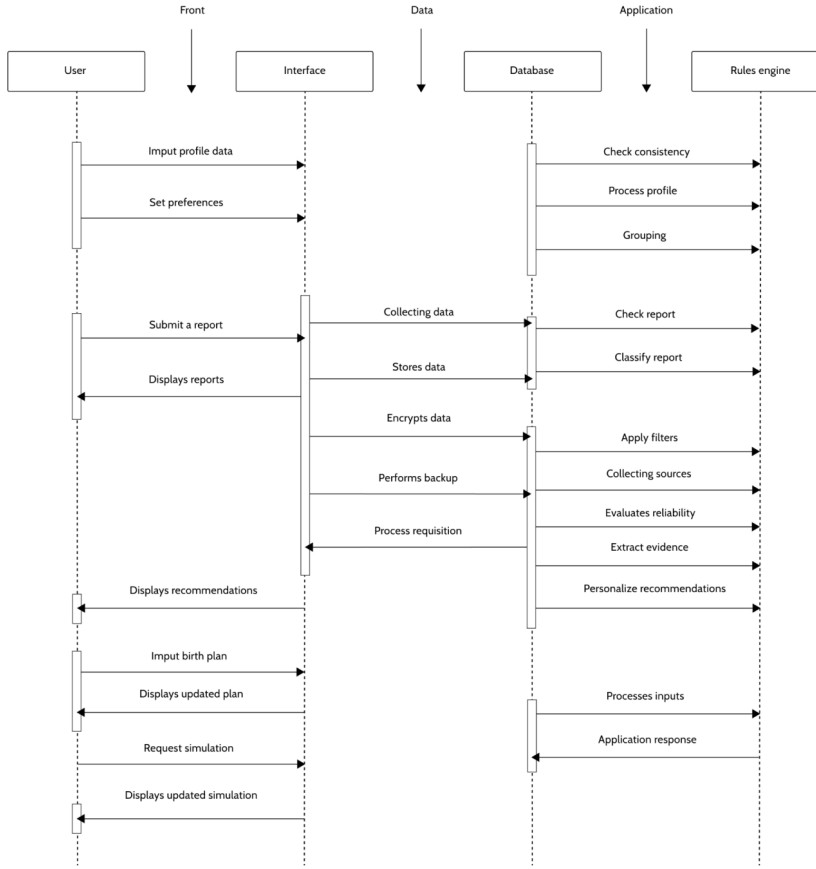


**Figure 3.** UML sequence diagram. **Source**. Author.

### 3.3.2 Main dataset

The formal description of data objects enables the capture of the logical structure that underlies the coherence of flows and the emergent behavior of the system. The decision to represent these components through three structural diagrams—class, object, and component—was made under the assumption that all functionality stems from structural entities (OMG, 2011).

#### *3.3.2.1 UML class diagram*

The class diagram illustrates the structural framework of the object-oriented system, delineating the fundamental classes, their attributes, and associated methods, as well as the composition, aggregation, and dependency links between them. The central class ProfileSegmentation encapsulates four entities: PersonalData, HistoryHealth, AccessCare, and Lifestyle, whose data are collected, validated, and encrypted before storage. The BirthPlan class is a comprehensive program designed to facilitate the aggregation of a pregnant woman's preferences, thereby enabling the generation and verification of personalized plans. The ObstetricInformationalBase is a comprehensive repository that consolidates profile data, birth plans, and evidences, thereby serving as a centralized information hub for the system. SimulationScenarios are models of decision scenarios. CuratingExperiences and PersonalizationRecommendations are concerned with empirical accounts and individualized recommendations, respectively. The data flow is enabled by the ProcessingData and Preprocessing classes, which are responsible for critical operations such as normalization, backup, linguistic adaptation, and pre-conditional verification. This modular, cohesive, and extensible structure facilitates the application's maintenance and evolution (Figure 4).

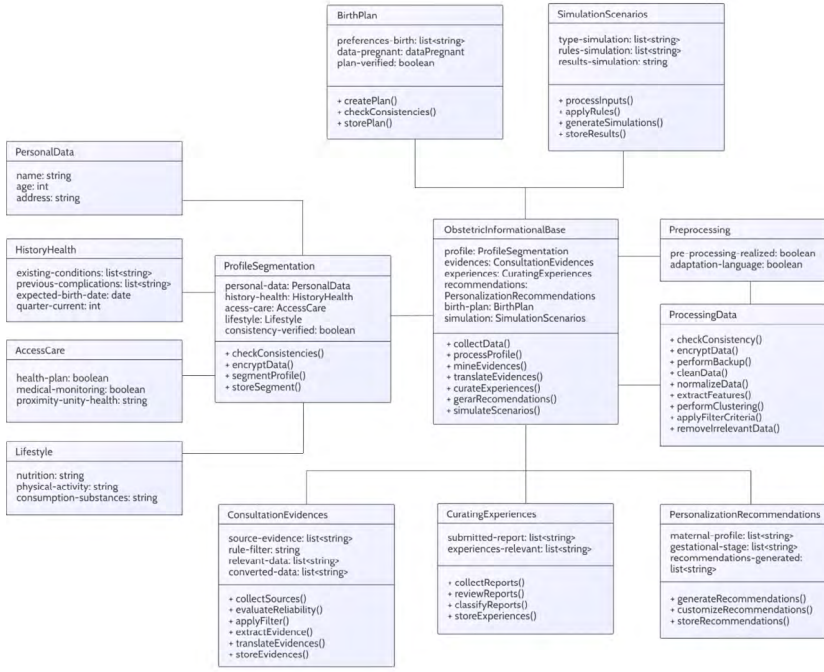


**Figure 4.** UML class diagram. **Source**. Author.

### *3.3.2.2 UML object diagram*

The object diagram instantiates the previously defined classes, thereby enabling the observation of the system in a real execution state. Each class is represented by an object that contains values derived from user input. For instance, the Patient Profile section includes specific data elements such as the patient's name, age, and clinical history. In contrast, the Birth Plan component reflects validated and adapted preferences. Objects such as Data Processing and Personalization of Recommendations operate with transformed and filtered data, delivering specific recommendations. In the Sources layer, clinical evidence and shared experiences are instantiated as Evidence Consultation and Curating Experiences, containing real data, such as scientific

articles and birth narratives. The model provides a precise view of the operational interdependencies and dynamic data flow between the system's objects (Figure 5).

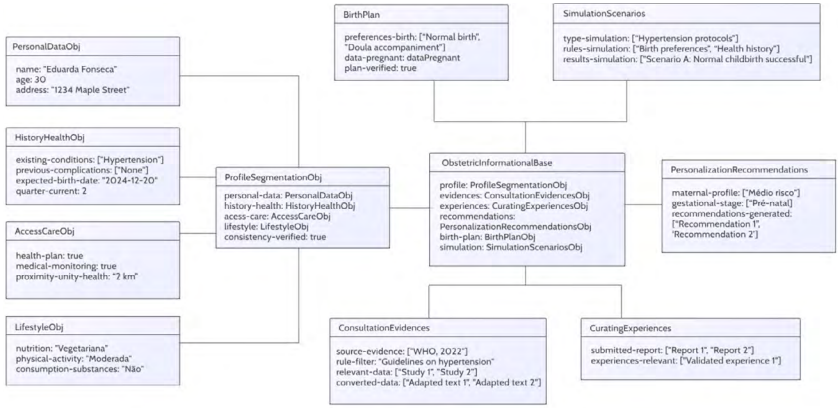


**Figure 5.** UML object diagram. **Source**. Author.

### *3.3.2.3 UML component diagram*

The following diagram illustrates the application's modular architecture, which is structured into three layers: User, Data Transformation, and Sources. The first category comprises interface components, including Patient Profile and Birth Plan, which are responsible for data collection and submission. These data are processed in the intermediate layer by modules such as Data Processing and Personalization of Recommendations, which perform normalizations, groupings, and guidance generation. The Scenario Simulation component utilizes these data to assess potential outcomes. Within the Sources layer, the Evidence Consultation and Curating Experiences modules facilitate the access and organization of external content, which subsequently feeds back into the system. Each component is connected by explicit interfaces such as InputProfile, OutputVerifiedPlan, and AccessRecommendations, ensuring data integrity, interoperability, and traceability. The diagram offers a clear visualization of the technical responsibilities of each functional block and the

relationships between them, facilitating the solution's maintenance and scalability (Figure 6).

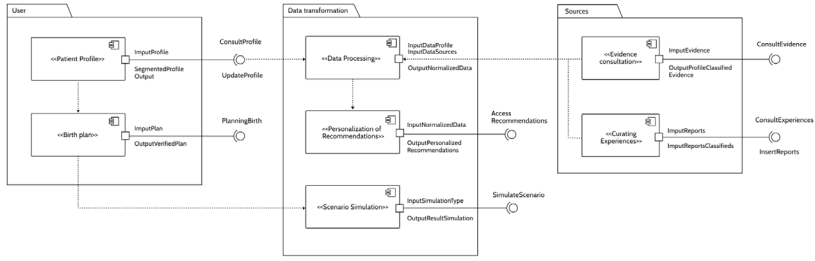


**Figure 6.** UML component diagram. **Source**. Author.

## 4 DISCUSSION

This study reflects the intersection between instances that are usually treated in a compartmentalized manner. It simultaneously articulates the three essential layers for its functioning: the health layer, anchored in clinical and biomedical guidelines; the informational layer, which structures data flow and its conversion into applicable knowledge; and the technological layer, which operationalizes these processes through digital solutions of algorithmic personalization. The model is predicated on the intricacy of contemporary obstetric care, endeavoring to transcend the mere dissemination of information and instead comprise a dynamic decision support ecosystem. In this ecosystem, scientific information is augmented by curation, with the objective of customizing guidance and interventions. The decision is founded on the objective of operationalizing the concept of patient-centered care within the paradigm of obstetric digital health, with a particular emphasis on empowering pregnant individuals and enhancing the quality of care. From this perspective, the multilayer system concept also allows for the articulation of the pregnant woman's local needs (specific, contextual, and individual) with the global requirements of best obstetric practices, based on scientific sources. The plurality applied here leverages the ability to

combine layers of micro- and macro-knowledge, optimizing the operationalization of personalized and evidence-based solutions within the same digital environment. Additionally, the concept of layers in health science alludes to the stratified structures of care needs—ranging from basic to highly specialized—and of knowledge dissemination itself, where raw data are transformed into information, and information, once processed and applied, becomes usable knowledge.

Another theoretical underpinning of this system is the integration of science and experience, grounded in the principles of practical epistemology (Sbissa et al., 2011) and evidence-based practice. This integration is further enriched by the incorporation of users' individual voices and narratives, a crucial element in transcending the conventional barriers encountered by health systems, which frequently exhibit a disconnect with contextual realities. The proposed system integrates these two levels, offering a balance between generalizable clinical recommendations and adaptation to each pregnant woman. This intersection is critical because it acknowledges that scientific knowledge, while robust, must always be contextualized by experience. This contextualization articulates subjective experiences with the universality of scientific evidence, thereby creating a dialogue between science and practice. The science layer ensures that recommendations are based on guidelines, while the experience layer adds the dimension of reality, promoting an interface where users can identify with similar stories and better assimilate recommendations.

Evidence-based recommendations, exemplified by the guidelines established by the WHO, serve as the foundation for regulatory frameworks that delineate care standards. These standards, in turn, function as a benchmark for the architecture of digital systems implemented by various nations. Conversely, the dissemination of these guidelines in an essentially textual format poses a series of challenges to their transposition into the social and digital environment, by requiring translation processes that, not infrequently, give rise to subjective interpretations by implementers and users. This subjectivity can compromise the uniformity of systems, which can lead to concrete risks of compromising content integrity. Such compromises can result in possible functional inconsistencies and limitations in the ability to audit technical and clinical compliance. This scenario has the potential to result in undesirable clinical outcomes and other adverse effects. In this

context, the transition from paper-based models to digital systems necessitates data standardization and the implementation of instruments to ensure technical interoperability, as well as the integrity and reliability of content integrated into systems. This mitigates risks associated with variability in guideline interpretation and safeguards the safety and effectiveness of healthcare.

The integration of formal modeling practices, particularly those facilitated by the UML, signifies a significant advancement in the conversion of technical-normative recommendations into operational artifacts. This integration provides a clear, auditable, and replicable computational structure, thereby enhancing the reliability and reproducibility of the research process. This conversion is not merely instrumental; it represents an effort of semantic transposition between clinical, informational, and computational domains. This transposition requires a multidimensional understanding of the interfaces between digital health, requirements engineering, and epistemology applied to care. The utilization of UML diagrams, encompassing use case, activity, sequence, class, object, and component, functions as a methodological axis to translate, with logical precision, the interaction flows between users, data, and decision engines. The model proposed here diverges from generic prescriptive approaches in that it is predicated on the articulation between the ontology of obstetric care and the algorithmic logic of informational personalization. This approach contributes to the modeling of a functional system and the construction of a normative-operational paradigm, in which the technical interface does not silence, but rather enhances the centrality of the gestational experience.

The technical dimension of the diagrams proved to be of particular strategic importance in delineating architectures oriented towards flow customization and profile segmentation. These stages, in turn, underpin information personalization and system responsiveness. The graphical materialization of components and processes ensures a formal representation that can be replicated, audited, and adapted to different institutional and cultural contexts. This mitigates the risks of interpretative ambiguity characteristic of exclusively textual guidelines. This assertion is particularly salient in light of the persistent challenges associated with the variability in the implementation of optimal obstetric practices, as highlighted by Kruk et al. (2018), and the opacity of numerous digital health systems with regard to the traceability

of decisions. The integration of an algorithmic layer that curates experiences and evidence establishes a hybrid knowledge framework that transcends technical rationality, thereby encompassing elements of practical epistemology and empathic listening. This integration of scientific evidence and subjective experience serves to revive the principle of informed autonomy, thereby transforming the care model from a hierarchical vertical structure to a dialogical paradigm that is responsive to individual needs. In this sense, the system proposes a technical–communicational mediation between biomedical knowledge and users' situated knowledge, reinforcing the notion of care as coproduction.

Nevertheless, the proposed modeling is not without limitations. Although UML allows for a robust formalization of systemic functionalities and flows, the system's real effectiveness depends on its validation in usage contexts. In such contexts, factors such as usability, health literacy, accessibility, and technological infrastructure can substantially interfere with the solution's performance. The absence of empirical validation, therefore, constitutes a relevant methodological limitation, which points to the need for future implementation studies, with an emphasis on analyzing user experience, clinical impact, and adherence to interoperability and information security guidelines (WHO, 2021). It is noteworthy that the proposed structure functions as a prospective catalyst for the reconfiguration of informational systems in maternal health. This reconfiguration involves the transformation of technical guidelines into interoperable computational architectures, with a focus on informational equity and the autonomy of pregnant women. The contribution of this study lies in demonstrating that diagrammatic modeling is not merely an illustrative supplement, but rather a structuring nucleus of an informational ecosystem that aims to enhance care quality through organizational intelligence and the technical translation of fundamental rights.

## 5 CONCLUSION

The analysis developed in this study highlights a significant discrepancy between the ideal design of obstetric care and its effective delivery. This discrepancy is evident in the practices of overuse, underuse, and inadequate use of resources, which persist

regardless of age categories, socioeconomic status, or health coverage configurations. Confronted with this structural predicament, which exposes inequalities and inadequacies in service delivery (Institute of Medicine US, 2001), there emerges a pressing need for instruments that overcome technological reductionism and digital determinism. In this sense, this study was dedicated to the conception and modeling of a multilayer Obstetric Informational Base, delineating a comprehensive proposal to address informational and humanization challenges in maternal care. The investigation's fundamental premise entailed the conceptual integration of two key elements: the humanization of obstetric care and the notion of informational empowerment. The overarching objective of this integration was twofold: first, to address and mitigate existing informational asymmetries within the healthcare system, and second, to empower pregnant individuals to become active participants in the decision-making process concerning their own healthcare. The adopted multilayer model, with its dynamic and static layers, reflects an intrinsic coherence with international guidelines for humanized health (WHO, 2015), aiming for adaptive care centered on individual needs.

The integration of ICTs and health was configured as a structuring axis of the proposal, conceived to overcome informational gaps and promote obstetric care based on human rights (WHO, 2012), democratizing access to validated scientific knowledge. The modeling diagrams developed for this proposal are instrumental in translating scientific knowledge into operational structures. These diagrams represent use cases, activities, sequences, classes, objects, and components. This diagrammatic approach provides a rigorous analysis and clear communication of the system's complexity, elucidating how informational empowerment and care personalization can be achieved and how each element contributes to a cohesive informational ecosystem. The objective of this modeling proposal is twofold: first, to contribute to the reduction of inappropriate obstetric practices, and second, to align with the recommendations of the WHO (2014, 2018) for a positive childbirth experience and the prevention of abuse. The conceptual model under discussion fosters care that is more respectful, safe, and values the pregnant woman's autonomy by grounding the provision of accessible and qualified information. Consequently, the modeled Obstetric Informational Base functions as a catalyst for transformation, promoting a culture of care grounded in

evidence and respect for human rights, in accordance with the biopsychosocial conception of health (WHO, 1946).

The project's primary objective was met through the development of a framework for an informational environment tailored to the needs of a particular audience within a complex ecosystem. This initiative aimed to address the biopsychosocial gap that hinders humanized approaches and the full exercise of citizenship within health, contributing to the field's advancement. It is acknowledged that, while the modeling of this platform signifies a significant advancement, its maximum impact on the comprehensive enhancement of obstetric health is contingent upon its strategic integration with other technologies and quality improvement initiatives, thereby constituting a multimodal intervention. The proposal's alignment with the 2030 Agenda (UN, 2015) and the SDGs, particularly SDG 3, is evident, as it aims to empower pregnant women with the knowledge necessary to manage their health. This study offers significant contributions to the fields of health management and public health. It proposes an innovative approach to restructuring obstetric care through the use of conceptual and diagrammatic modeling. This proposal delineates a range of opportunities for enhancing the quality of care, expanding the array of safe, effective, and equitable interventions, and shifting towards obstetric care that is more humane and centered on women's autonomy.


## *Funding*

This study received support from the National Council for Scientific and Technological Development (CNPq) and the Federal University of Santa Catarina (UFSC), which supported this study in part.


## *Conflict of interest*

The authors declare that there are no conflicts of interest related to this research.

### *Contribution statement*

**Conceptualization:** Paulianne Fontoura Guilherme de Souza, Douglas Dyllon Jeronimo de Macedo
**Methodology:** Paulianne Fontoura Guilherme de Souza, Douglas Dyllon Jeronimo de Macedo
**Data Curation:** Paulianne Fontoura Guilherme de Souza
**Formal Analysis:** Paulianne Fontoura Guilherme de Souza
**Writing – Original Draft:** Paulianne Fontoura Guilherme de Souza, Gustavo Generaldo de Sá Teles Junior, Douglas Dyllon Jeronimo de Macedo
**Writing – Review & Editing:** Paulianne Fontoura Guilherme de Souza, Gustavo Generaldo de Sá Teles Junior, Douglas Dyllon Jeronimo de Macedo
**Supervision:** Douglas Dyllon Jeronimo de Macedo

### *Statement of data consent*

The bibliographic data used in this study, including data from Web of Science and PubMed, have been processed and analyzed as described in "Methodology." The datasets generated during this research are available upon request and can be provided to interested researchers for further review.

## REFERENCES


Almeida Junior, O. F. (2009). Mediação da informação e múltiplas linguagens. *Tendências da Pesquisa Brasileira em Ciência da Informação*, *2*(1), 89–101. http://hdl.handle.net/20.500.11959/brapci/119300

Araújo, E. A. de. (1992). Informação, cidadania e sociedade no Brasil. *Informação & Sociedade: Estudos*, *2*(1), 42–49.

Brasil. (2005, 8 de abril). Lei nº 11.108, de 7 de abril de 2005. Altera a Lei nº 8.080, de 19 de setembro de 1990, para garantir às parturientes o direito à presença de acompanhante durante o trabalho de parto, parto e pós-parto imediato, no âmbito do Sistema Único de Saúde—SUS. Diário Oficial da União, Seção 1, p. 1. https://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/lei/l11108.html

Brasil. Ministério da Saúde. (2011, 24 de junho). Portaria GM/MS nº 1.459, de 24 de junho de 2011. Institui, no âmbito do Sistema Único de Saúde — SUS — a Rede Cegonha. Diário Oficial da União. https://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt1459_24_06_2011.html

Brizuela, V., & Tunçalp, Ö. (2017). Global initiatives in maternal and newborn health. *Obstetric Medicine*, *10*(1), 21–25. https://doi.org/10.1177/1753495X16684987

Coulm, B., Le Ray, C., Lelong, N., Drewniak, N., Blondel, B., & Zeitlin, J. (2012). Obstetric interventions for low-risk pregnant women in France: Do maternity unit characteristics make a difference? *Birth*, *39*(3), 183–191. https://doi.org/10.1111/j.1523-536X.2012.00547.x

Diniz, S. G. (2009). Gênero, saúde materna e o paradoxo perinatal. *Revista Brasileira de Crescimento e Desenvolvimento Humano*, *19*(2), 313–326. https://doi.org/10.7322/jhgd.19944

D'Oliveira, A. F. P. L., Diniz, S. G., & Schraiber, L. B. (2002). Violence against women in health-care institutions: An emerging problem. *The Lancet*, *359*(9318), 1681–1685. https://doi.org/10.1016/S0140-6736(02)08592-6

Domingues, R. M. S. M., Santos, E. M. dos, & Leal, M. da C. (2004). Aspectos da satisfação das mulheres com a assistência ao parto: Contribuição para o debate. *Cadernos de Saúde Pública*, *20*(Suppl. 1), S52–S62. https://doi.org/10.1590/s0102-311x2004000700006

Donabedian, A. (2005). Evaluating the quality of medical care. *The Milbank Quarterly*, *83*(4), 691–729. (Reprinted from *The Milbank Memorial Fund Quarterly*, *44*(3, Pt. 2), pp. 166–203, 1966). https://doi.org/10.1111/j.1468-0009.2005.00397.x

Downe, S., Finlayson, K., & Fleming, A. (2015). What matters to women: A systematic scoping review to identify the processes and outcomes of antenatal care provision that are important to healthy pregnant women. *BJOG: An International Journal of Obstetrics & Gynaecology*, *123*(4), 529–539. https://doi.org/10.1111/1471-0528.13847

EURO-PERISTAT Project with SCPE and EUROCAT. (2013, May). European Perinatal Health Report: The health and care of pregnant women and babies in Europe in 2010. https://www.europeristat.com/images/European%20Perinatal%20Health%20Report_2010.pdf

Fuck, M. P., & Vilha, A. M. (2012). Inovação tecnológica: Da definição à ação. *Contemporâneos: Revista de Artes e Humanidades*, (9), 1–17.

Guimarães, M. C., & Silva, C. H. (2011). Acesso à informação em saúde: Por uma agenda política. In *Anais do XII Encontro Nacional de Pesquisa em Ciência da Informação* (pp. 3551–3563). UNB; ANCIB. https://www.arca.fiocruz.br/handle/icict/327

Hamilton, K., & Miles, R. (2006). *Learning UML 2.0*. O'Reilly. https://jti.polinema.ac.id/wp-content/uploads/2019/02/Buku-Learning-UML-2.0.pdf

Institute of Medicine (US) Committee on Quality of Health Care in America. (2001). *Crossing the quality chasm: A new health system for the 21st century*. National Academies Press.

Kruk, M. E., Gage, A. D., Arsenault, C., Jordan, K., Leslie, H. H., Roder-DeWan, S., Adeyi, O., Barker, P., Daelmans, B., Doubova, S. V., English, M., García-Elorrio, E., Guanais, F., Gureje, O., Hirschhorn, L. R., Jiang, L., Kelley, E., Lemango, E. T., Liljestr, J., … Pate, M. (2018). High-quality health systems in the Sustainable Development Goals era: Time for a revolution. *The Lancet Global Health*, *6*(11), e1196–e1252. *https://doi.org/10.1016/S2214-109X(18)30386-3*

Leite, R. A. F., Brito, E. S., Silva, L. M. C., Palha, P. F., & Ventura, C. A. A. (2014). Access to healthcare information and comprehensive care: Perceptions of users of a public service. *Interface (Botucatu)*, *18*(51), 661–671. https://doi.org/10.1590/1807-57622013.0334

Leite, T. H., Marques, E. S., Esteves-Pereira, A. P, Nucci, M. F., Santos, Y. R. P, & Leal, M. C. (2021). Desrespeitos e abusos, maus tratos e violência obstétrica: um desafio para a epidemiologia e a saúde pública no Brasil. *Ciência & Saúde Coletiva*, *27*(2), 483–491. *https://doi.org/10.1590/1413-81232022272.38592020*

Macedo, D. D. J., Von Wangenheim, A., & Dantas, M. A. R. (2015). A data storage approach for large-scale distributed medical systems. In *2015 Ninth International Conference on Complex, Intelligent, and Software Intensive Systems* (pp. 486–490). Santa Catarina, Brazil. *https://doi.org/10.1109/CISIS.2015.88*

Ministério da Saúde. (2000). *Humanização do parto: Humanização no pré-natal e nascimento*.

Mota, F. R. L., Araujo, N. C., & Santos, P. A. I. C. (2015). Necessidades informacionais das gestantes atendidas em unidades básicas de saúde do bairro Benedito Bentes—Maceió/AL [Paper presentation]. In *XVI Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB)*. João Pessoa, PB, Brasil. *http://www.ufpb.br/evento/index.php/enancib2015/enancib2015/paper/viewFile/2962/1267*

Object Management Group (OMG). (2011, August). *Unified modeling language TM: Superstructure (Version 2.4.1)*. *https://www.omg.org/spec/UML/2.4.1/Superstructure/PDF*

Office of the United Nations High Commissioner for Human Rights (OHCHR). (2012). *Technical guidance on the application of a human rights-based approach to the implementation of policies and programmes to reduce preventable maternal morbidity and mortality (A/HRC/21/22)*. *https://www2.ohchr.org/english/issues/women/docs/A.HRC.21.22_en.pdf*

Pan American Health Organization (PAHO). (2019). *Recomendaciones de la OMS: Cuidados durante el parto para una experiencia de parto positiva*.

Puel, A., Wangenheim, A. V., Meurer, M. I., & Macedo, D. D. J. (2014). BUCOMAX: Collaborative multimedia platform for real time manipulation and visualization of bucomaxillofacial diagnostic images. In *2014 IEEE 27th International Symposium on Computer-Based Medical Systems* (pp. 392–395). New York, NY, USA. *https://doi.org/10.1109/CBMS.2014.12*

Reneker, M. H. (1993). A qualitative study of information seeking among members of an academic community: Methodological issues and problems. *Library Quarterly*, *63*(4), 487–507. *https://doi.org/10.1086/602618*

Renfrew, M. J., McFadden, A., Bastos, M. H., Campbell, J., Channon, A. A., Cheung, N. F., Silva, D. R. A. D., Downe, S., Kennedy, H. P., Malata, A., McCormick, F., Wick, L., & Declercq, E. (2014). Midwifery and quality care: Findings from a new evidence-informed framework for maternal and newborn care. *The Lancet*, *384*(9948), 1129–1145. *https://doi.org/10.1016/S0140-6736(14)60789-3*

Royal College of Obstetricians and Gynaecologists. (2017). *Annual review 2016/2017*. *https://www.rcog.org.uk/media/thmfgugt/2017.pdf*

Russo, J. A., & Carrara, S. L. (2015). Sobre as ciências sociais na saúde coletiva – com especial referência à antropologia. *Physis: Revista de Saúde Coletiva, 25*(2), 467–484. *https://doi.org/10.1590/S0103-73312015000200008*

Sbissa, P. P. M., Schneider, D. R., & Sbissa, A. S. (2011). Characterization of the epistemological development of health and complementary practices. *Arquivos Catarinenses de Medicina, 40*(2), 70–77. *https://www.acm.org.br/revista/pdf/artigos/871.pdf*

Silva, E. L., & Menezes, E. M. (2001). *Metodologia da pesquisa e elaboração de dissertação* (3ª ed. rev. atual.). Laboratório de Ensino a Distância da UFSC.

Soares, T. S., Dantas, M. A. R., de Macedo, D. D. J., & Bauer, M. A. (2013). A data management in a private cloud storage environment utilizing high performance distributed file systems. In *2013 Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises* (pp. 158-163). Hammamet, Tunisia. *https://doi.org/10.1109/WETICE.2013.12*

Souza, V. B., Roecker, S., & Marcon, S. S. (2011). Ações educativas durante a assistência pré-natal: Percepção de gestantes atendidas na rede básica de Maringá-PR. *Revista Eletrônica de Enfermagem, 13*(2), 199–210. *https://doi.org/10.5216/ree.v13i2.10450*

Souza Inácio, A. de, Andrade, R., von Wangenheim, A., & de Macedo, D. D. J. (2014). Designing an information retrieval system for the STT/SC. In *2014 IEEE 16th international conference on e-health networking, applications and services (Healthcom)* (pp. 500–505). Natal, Brazil. *https://doi.org/10.1109/HealthCom.2014.7001893*

Tamrat, T., Ratanaprayul, N., Barreix, M., Tunçalp, O., Lowrance, D., Thompson, J., Rosenblum, L., Kidula, N., Chahar, R., Gaffield, M. E., Festin, M., Kiarie, J., Taliesin, B., Leitner, C., Wong, S., Wi, T., Kipruto, H., Adegboyeg, A., Muneene, D., Say, L., & Mehl, G. (2022). Transitioning to digital systems: The role of World Health Organization's Digital Adaptation Kits in operationalizing recommendations and interoperability standards. *Global Health: Science and Practice, 10*(1), Article e2100320. *https://doi.org/10.9745/GHSP-D-21-00320*

Targino, M. das G. (1991). Biblioteconomia, informação e cidadania. *Revista da Escola de Biblioteconomia da UFMG*, *20*(2), 149–160. *https://www.brapci.inf.br/index.php/article/download/13455*

Tobar, F., & Romano Yalour, M. (2001). *Como fazer teses em saúde pública*. Fiocruz.

Triviños, A. N. S. (1987). *Introdução à pesquisa em Ciências Sociais: A pesquisa qualitativa em Educação*. Atlas.

United Nations (UN). (2015). *Transforming our world: The 2030 Agenda for Sustainable Development*. *https://sdgs.un.org/2030agenda*

World Health Organization (WHO). (1946). *Constitution of the World Health Organization*. *https://www.who.int/about/governance/constitution*

World Health Organization (WHO). (2014). *Prevenção e eliminação de abusos, desrespeito e maus-tratos durante o parto em instituições de saúde*. *https://iris.who.int/bitstream/handle/10665/134588/WHO_RHR_14.23_por.pdf*

World Health Organization (WHO). (2015). *Global strategy for women's, children's and adolescents' health (2016-2030)*. *https://www.who.int/publications/i/item/9789241510348*

World Health Organization (WHO). (2018). *WHO recommendations: Intrapartum care for a positive childbirth experience*. *https://iris.who.int/bitstream/handle/10665/260178/9789241550215-eng.pdf?sequence=1&isAllowed=y*

World Health Organization (WHO). (2019). *WHO guideline: Recommendations on digital interventions for health system strengthening*. *https://iris.who.int/bitstream/handle/10665/311941/9789241550505-eng.pdf?sequence=31*

World Health Organization (WHO). (2021). *Digital adaptation kit for antenatal care: Operational requirements for implementing WHO recommendations in digital systems*. *https://iris.who.int/bitstream/handle/10665/339745/9789240020306-eng.pdf?sequence=1*

World Health Organization (WHO), & International Telecommunication Union (ITU). (2012). *National eHealth strategy toolkit*.

Zorzam, B. A. O. Z. (2013). *Informação e escolhas no parto: Perspectivas das mulheres usuárias do SUS e da saúde suplementar* [Dissertação de Mestrado, Universidade de São Paulo]. Repositório de Teses e Dissertações da USP. *https://www.teses.usp.br/teses/disponiveis/6/6136/tde-10112013-223016/publico/BiancaAlves.pdf*

CHAPTER 3

# DATA PROVENANCE AND BLOCKCHAIN: AN APPROACH IN THE CONTEXT OF HEALTH INFORMATION SYSTEMS

**Márcio José Sembay**
*Department of Information Science, Federal University of Santa Catarina, Brazil.*
ORCID: *https://orcid.org/0000-0002-7648-8861*
**Email:** *marcio.sembay@posgrad.ufsc.br*

**Douglas Dyllon Jeronimo de Macedo**
*Department of Information Science, Federal University of Santa Catarina, Brazil.*
ORCID: *https://orcid.org/0000-0002-3237-4168*
**Email:** *douglas.macedo@ufsc.br*

**Alexandre Augusto Gimenes Marquez Filho**
*Integrated Telemedicine and Telehealth System of Santa Catarina, Federal University of Santa Catarina, Brazil.*
ORCID: *https://orcid.org/0000-0002-2656-9479*
**Email:** *alexandre.agmf@gmail.com*

## ABSTRACT

The integration of data provenance and blockchain, in accordance with international health standards, was demonstrated to enhance patient data management through seamless integration with health information systems (HIS). This study built upon the findings of previous research conducted by the same authors, with the objective of conducting a more comprehensive and in-depth

analysis. In terms of methodology, this research was a basic study characterized as a bibliographical and exploratory investigation with a qualitative approach. The analyses carried out, based on related work, focused on the relationships between the main applications of data provenance in conjunction with the intrinsic characteristics of blockchain technology. These aspects were examined in the context of HIS, which made it possible to identify the international data interoperability standards specifically adopted in electronic health records (EHRS) and personal health records (PHRS). The primary outcomes of this study included the identification of the relationships between the primary applications of data provenance and the characteristics of blockchain, with a particular focus on HIS. Additionally, the analysis of the literature on data provenance and blockchain technology led to the recognition of the main interoperability standards. This culminated in a reflective synthesis of the findings. A comprehensive analysis of the results, grounded in the identified fundamental elements, yielded significant insights into the integration of data provenance and blockchain technology within the HIS, particularly in the context of EHR and PHR.

**KEYWORDS**: data provenance, blockchain, health information systems, electronic health record, personal health record

**HOW TO CITE**: Sembay, M. J., Dyllon Jeronimo de Macedo, D., & Augusto Gimenes Marquez Filho, A. (2025). Data provenance and blockchain: An approach in the context of health information systems. In A. Semeler (Ed.), Artificial Intelligence and Data Science Practices in Scientific Development, *Advanced Notes in Information Science, volume 8* (pp. 73-121). Pro-Metrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9331-4-5.111.

**COPYRIGHT**: © 2025 The author(s). This article is distributed under the terms of the CC BY-NC 4.0 license, which permits copying and redistribution of the material in any medium or format, adaptation, transformation, and building upon the material, provided that the license terms are followed.

## 1 INTRODUCTION

The health sector has been identified as a primary beneficiary of communication through information systems (IS) and information and communication technologies (ICT). These technologies have been found to support and record actions in the health context, encompassing operational, managerial, and decision support functions (World Health Organization, 2008). Consequently, the integration of ICT by competent health professionals is hypothesized to facilitate the enhancement of national health systems (Weerakoon & Chandrasiri, 2023). In this scenario, it is understood that such technologies enable the consolidation of a technological ecosystem focused on promoting human health, playing a central role in the digital transformation of care systems and the personalization of health services, as in the case of Health Information Systems (HIS), defined as "data, information, and knowledge processing systems in healthcare environments" (Haux, 2006). The global health sector is marked by an escalating volume of data pertaining to patient care requirements (Dash et al., 2019). This augmentation in data production encompasses hospital records, examination results, devices that are part of the Internet of Things (IoT), and other medical data (Dash et al., 2019). At present, we are confronted with an immense inundation of data pertaining to a myriad of aspects of life, with a particular emphasis on the healthcare sector. As in other fields, healthcare organizations have been producing data at an accelerated pace, which brings both significant benefits and challenges. Technological advances have led to exponential data generation, which has made its management a complex task, especially when using conventional technologies. This complexity is further compounded in the context of IoT devices, whose structures are guided by user-centered design (Dash et al., 2019; Samuel & Garcia-Constantino, 2022).

In this context, the growing demand for managing large volumes of data in HIS has led to the adoption of computational strategies that enable the historical processing of this information. Examples of such strategies include data provenance and the use of blockchain. The primary application of blockchain technology is tracking provenance, as it provides robust mechanisms to ensure the integrity and security of databases associated with provenance information (Greenspan, 2016). Data provenance is

a critical process for providing a comprehensive view of the data utilized in IS, with an emphasis on identifying its origins and the transformations it has undergone over time. This approach has been applied in various computational contexts, with a particular emphasis on the health domain (Sembay et al., 2020). In recent years, there has been a notable increase in the application of data provenance in scientific research focused on health-related fields, encompassing a diverse array of experiments. The technologies employed in this domain have demonstrated substantial and encouraging outcomes (Sembay et al., 2021). In this scenario, data provenance establishes itself as a fundamental foundation for ensuring the quality of medical data, as well as for strengthening the protection of patient privacy (Margheri et al., 2020).

In the domain of healthcare, blockchain technology has emerged as a reliable and consensus-based distributed ledger solution, enabling the development of interoperable, auditable, and secure systems (Swan, 2015). In the healthcare sector, its implementation entails the management of access to and dissemination of sensitive data, the enhancement of service transparency and auditability, and the assurance of data interoperability, among other pivotal applications (Monteil, 2019). Moreover, acknowledging the identified lacuna in the extant literature concerning the integration of data provenance and blockchain in HIS, this study endeavored to address three fundamental inquiries: (1) What are the conceptual and practical relationships between data provenance and blockchain technologies? (2) To what extent can the integration of data provenance mechanisms with blockchain contribute to the effectiveness, security, and interoperability of HIS? and (3) What types of data interoperability patterns can be observed in the joint use of data provenance and blockchain in HIS? These inquiries were addressed through an analytical process encompassing a theoretical review and an evaluation of practical applications within the framework of HIS. The objective of this article is to extend the research of Sembay et al. (2022). Sembay et al. conducted a study on the combined use of data provenance technologies and models and blockchain technologies employed in HIS, specifically in electronic health record (EHR) and personal health record (PHR).

It is hypothesized that this study will facilitate a more comprehensive examination of the preceding study by Sembay et al. (2022) on the primary applications of data provenance, as

delineated in the research of Simmhan et al. (2005), in conjunction with the characteristics of the blockchain as expounded by Sultan et al. (2018). This examination aims to ascertain the potential relationships between these two technological entities. Furthermore, the analysis was expanded in relation to the related work presented by Sembay et al. (2022), which addresses the joint application of data provenance and blockchain in the context of EHR and PHR. This expansion facilitated a more profound comprehension of the subject matter and enabled the identification of the predominant health data interoperability standards that have been adopted in these studies. To complement this expansion of the analysis initially developed by Sembay et al. (2022), an analytical synthesis was drawn up that makes it possible to reflect on the relevance of the elements identified, further broadening the understanding of the integrated use of data provenance and blockchain in EHR and PHR systems. Concurrently, the extant literature on the subject was expanded, thereby providing a more comprehensive basis for understanding the topic. Subsequent to this introduction, the literature review is presented, followed by the outline of the methodological approach. The results of the study are subsequently presented, summarized, and discussed. The paper concludes with a summary of its key points and a list of references.

### *1.1 Literature review*

The literature review examines the concepts of HIS and data interoperability, with a particular emphasis on data provenance and blockchain technology. The text undertakes an examination of the primary provenance models, their applications in the health context, and their integration with standards such as Health Level 7 (HL7) Fast Healthcare Interoperability Resource (FHIR) and World Wide Web (W3C) PROV. The role of blockchain in HIS interoperability is also presented, highlighting its characteristics and applications. Consequently, studies integrating data provenance and blockchain in HIS scenarios are presented.

### 1.1.1 Health information systems

Health information systems are comprehensive platforms designed to collect, process, communicate, and utilize essential health data to enhance the efficiency and effectiveness of healthcare services. These systems play a crucial role in supporting management and decision-making across all levels of the healthcare sector. Health information systems are being increasingly adopted in various domains, ranging from administrative functions to clinical decision support (Sembay & Macedo, 2022). By generating high-quality and relevant information, they contribute significantly to the planning, execution, and evaluation of health programs (Haux, 2006; World Health Organization, 2004). Health information systems has been increasingly adopted across the globe to enhance hospital efficiency, the quality of service, and patient satisfaction (Cesnik & Kidd, 2010). They can also be regarded as a system of information, integrating the collection, processing, communication, and utilization of critical information. The purpose of this integration is to improve the efficiency of health services by means of enhanced management in all health sectors. This system has been demonstrated to produce relevant information of superior quality to support the management and planning of health programs (Haux, 2006; World Health Organization, 2004). The broad categorization of HIS can be subdivided into two primary classifications: systems dedicated to the recording of individual-level health data, and systems focused on the aggregation of data for decision-making and information governance, which is colloquially referred to as health information management systems (Dehnavieh et al., 2018). It is imperative to underscore that HIS facilitate the digitalization of all patient-related information, thereby enhancing the quality and efficiency of healthcare delivery (Al Jarullah & El-Masri, 2012). In this regard, HIS are characterized as a computerized system for collecting, storing, and retrieving information concerning individuals involved in the healthcare domain—including patients, physicians, nurses, and other professionals responsible for generating clinical and administrative data. This process is executed across both local and national contexts, irrespective of whether the environments are integrated or distributed (Andargolia et al., 2017; Robertson et al., 2010; Sligo et al., 2017). Regarding this, we point out some of the main existing HIS, which are as follows:

1. Electronic health record (EHR): This refers to the concept of a comprehensive, interinstitutional, and longitudinal electronic record of patient health data. This type of record includes not only information directly related to medical assessment and treatment but also data relevant to an individual's overall health status (Hoerbst & Ammenwerth, 2010). It is imperative to acknowledge that discourse pertaining to EHRs frequently pertains to the Health Insurance Portability and Accountability Act (HIPAA), a US federal statute promulgated in 1996, initially conceived to safeguard health insurance coverage for employees and their dependents (Annas, 2003).

2. Personal health record (PHR): These are health records that are frequently created and managed by the patients themselves. They may be desktop-based, web-based, or accessible via mobile devices such as smartphones or portable storage units (Liu et al., 2011).

3. Learning health system (LHS): This is a system designed to collect, share, and utilize health data to rapidly generate knowledge and support transformative decision-making that contributes to improved health outcomes. The system's operational framework is characterized by its ability to adapt to varying demands, a capability facilitated by its integration of technology, processes, and policies (Friedman et al., 2015).

4. Healthcare monitoring system (HMS): This focuses on health monitoring through the application of wearable and environmental sensors. These sensors have been developed for the purpose of collecting health-related data in patients' or users' everyday environments (Korhonen et al., 2003).

5. Clinical research information system (CRIS): This is a software system designed to support clinical research. The primary objective of CRIS is to reduce the costs of scientific studies. CRIS integrates clinical care, research data collection, and support for hospital operations (Nadkarni et al., 2012).

6. Hospital information system (HIS): In this context, the HIS can be identified as a computerized information system installed in a hospital environment with the objective of recording patient information, thereby enabling its dissemination to all sectors of the hospital that require it. An HIS is designed to support multiple functionalities, including patient care management and hospital administration, covering six distinct purposes: patient management, department management, clinical documentation, clinical decision support, financial resource management, and healthcare manager support (Ismail et al., 2010).

7. Radiology information system (RIS): This emerged with the implementation of computers in hospitals, when it was recognized that they could be used as an aid in the field of radiology (Bakker, 1991). A RIS is a specialized software designed to facilitate the management of radiology departments. It enables the reception of interpretations and the generation of patient lists. This system has the capacity to generate historical reports from radiologists and frequently transmits the final report to the HIS (Honeyman, 1999).

8. Laboratory information system (LIS): This is defined as a set of interconnected software applications designed to manage information within a clinical analysis laboratory. These applications may address technical, operational, administrative, managerial, or a combination of these aspects, with the overarching objective being the effective management of data within the laboratory setting. It is imperative to conceptualize it as an entity independent of laboratory automation systems (LAS), with which it can establish a relationship of profound intimacy, bordering on symbiosis. However, for the purpose of ascertaining its true purpose, it is essential to disengage from these systems. Laboratory automation, in turn, can be conceptualized as a component of the LAS, a comprehensive framework encompassing the management of process activities involved in the oversight of laboratory equipment and instruments, sample control, and analytical processes (Blick, 1997).

9. Picture archiving and communication system (PACS): This system consists of interconnected subsystems that utilize computer networks for the acquisition, storage, and visualization of images and data. The complexity of these systems can range from a rudimentary integration with a modality and a visualization station, accompanied by a modest database, to a sophisticated system that oversees the management of medical images across a medium or large hospital (Zhang et al., 2003). In essence, Law and Zhou (2003) offer a concise definition of the PACS as an information technology system responsible for the transmission and storage of medical images. They assert that a PACS comprises interface components for HIS/RIS, imaging modalities such as digital imaging and communications in medicine (DICOM), storage control, and viewing stations.

These HIS are implemented in various countries and play a crucial role in managing numerous processes related to health data. By facilitating the aggregation, storage, dissemination, and examination of clinical and administrative data, these systems substantially enhance the efficacy, precision, and coherence of healthcare administration. The implementation of these tools has been demonstrated to facilitate enhanced decision-making processes, improve patient care, and promote interoperability among healthcare institutions.

### 1.1.2 Data interoperability in HIS

The necessity for interoperability standards between HIS is intrinsic to facilitate communication and exchange of health data, thereby establishing mechanisms for interoperability among disparate health platforms. Therefore, for HIS to fulfill their role, it is essential that they possess computational tools capable of executing and mediating the entire process of interoperability of health data. In this sense, some of the most used interoperability standards in different HIS are as follows:

1. DICOM: This was developed through a collaborative effort between the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA). DICOM is an object-oriented standard that defines

information objects, services, and classes of services that perform these services. Each device is equipped with a set of predetermined objects that are designed to recognize the file and facilitate access to it and the associated services. Additionally, these objects enable the negotiation process between two devices to determine which one should transfer the image. The DICOM standard has been adopted by medical equipment manufacturers and healthcare informatics systems developers as the standard for exchanging images in a digital format (Honeyman, 1999; Mildenberger et al., 2002; Oosterwijk, 2002).

2. HL7 FHIR: Achieving software interoperability in the healthcare domain is possible through the implementation of consistent standards, such as HL7, a standards development organization that facilitates the exchange, integration, sharing, and retrieval of healthcare information. In this regard, the FHIR created by HL7 is another significant standard that describes data formats and elements, as well as an application programming interface for interoperable EHR exchange. Consequently, HL7 FHIR has been established as a standard that defines resources, including content definitions, architecture, models, and paradigms for exchanging health information (HL7 International, n.d.).

3. Integration of the healthcare enterprise (IHE): It was initiated in November 1998. IHE is a high-level information model designed to facilitate adaptations to the HL7 and DICOM standards. The initial objective of IHE was to establish and promote the utilization of standards, with the aim of ensuring the compliance of equipment and IS. This initiative was designed to enhance the efficiency of daily clinical operations (Huang, 2019). Consequently, the IHE makes an effective contribution to all health professionals, who can signal the main instances that emerge daily in the range of vision of their activities. While originally specified for radiology, the current objective is to establish rules for identifying and resolving the challenges that hinder the effective and functional integration of HIS. This initiative involves collaboration with medical specialists and information technology professionals. The technical architecture of the

IHE delineates a common language, vocabulary, and model using DICOM and HL7 to complete a well-defined radiological suite and clinical transactions for specific services (Bernardini et al., 2003; Huang, 2019). The objective of the IHE is to furnish the end user with enhanced access to critical and clinical patient information stored in all systems connected to a hospital network. The overarching aim is to facilitate efficiency, prognosing and integrating functionalities between incompatible systems (Boochever, 2004).

4. Extract–transform–load (ETL): This refers to a widely used process for integrating data from multiple sources or applications, including those from different domains. Extraction, transformation, and loading constitute a data management method comprising three primary phases, with the objective of preparing data for operational or analytical use. The extracted data is typically loaded into a target database, such as a data warehouse, especially for operational analytics (Bansal, 2014). The following stages comprise the fundamental phases of the process: Extract: The initial phase of the process is defined as the extraction of data from relevant data sources. These sources may be in flat file formats such as (.csv), (.xls), and (.txt), or accessed via a RESTful client. Transform: During this stage, the extracted data are cleaned and converted to comply with the schema of the target database. Common transformation tasks include data normalization, duplicate removal, integrity constraint checks, filtering based on regular expressions, data sorting and grouping, and the application of built-in functions as needed. Load: The final phase of the process involves loading the transformed data into a data warehouse. This is typically done to support Big Data environments and large-scale data analysis (Bansal, 2014).

5. Cross enterprise document sharing (XDS): This addresses the need for the registration, distribution, and access across health enterprises of patients' clinical information (Noumeir & Renaud, 2010).

6. HL7 clinical document architecture (CDA): This is a set of guidelines that define the syntax rules and provide a

fundamental framework for implementing the semantics of a clinical document. This facilitates the electronic exchange of clinical documents (Dolin et al., 2001).

### 1.1.3 Data provenance

Data provenance, as defined by Buneman et al. (2001), refers to the complementary documentation associated with a specific dataset. This documentation captures information about how, when, and why the data were generated, as well as by whom. This metadata plays a crucial role in ensuring the quality, authenticity, and trustworthiness of data by enabling the identification of their origin, the detection of potential errors, and the attribution of data sources (Margheri et al., 2020). Additionally, data provenance can be defined as a set of descriptive records that trace the historical derivation of a data product from its original sources. It is widely recognized as a fundamental element for ensuring the reproducibility of results, facilitating data sharing, and promoting the reuse of knowledge within the scientific community (Freire et al., 2008). In addition to its pertinence in scientific research, data provenance has also gained significance in domains such as healthcare, finance, and artificial intelligence (AI). In these fields, transparency, traceability, and accountability are paramount for compliance, auditing, and ethical data use. A fundamental aspect of data provenance is causality, which pertains to the description of the process—along with its input data and parameters—that results in the creation of a final dataset. This component is responsible for the documentation of process dependencies, thereby facilitating both the reproduction and validation of data workflows. According to Freire et al. (2008), prospective provenance specifies the intended steps to generate a data product (e.g., processes, workflows, or scripts), while retrospective provenance captures the actual execution, including system settings, inputs, outputs, and runtime parameters. In summary, prospective provenance delineates the recommended course of action, while retrospective provenance documents the actions that have been executed. This provides a foundational framework for transparency and reproducibility.

In general terms, the operation of data provenance involves tracking the movement and transformation of data during the

execution of queries and programs. In the event of such operations, data are transferred from one database to another, and a description of the relationships and processes involved is generated (Tan, 2008). In this context, data provenance is a critical element, as it facilitates the tracking of data origins, the documentation of its trajectory across various sources, and the identification of transformations and dependencies (Simmhan et al., 2005). This tracking capability is imperative for ensuring data transparency, auditability, and reliability, particularly in complex data environments.

#### *1.1.3.1 Data provenance: Main models*

To ensure the successful provenance of data in a variety of application scenarios, the creation of models will be undertaken. Consequently, initiatives to represent provenance through informational resources in general commenced with discussions on the construction of the open provenance model (OPM) in 2006, at the first International Provenance and Annotation Workshop (IPAW) (Moreau, 2006). The proposal of OPM was to define a data model that is open from an interoperability point of view, but also with respect to the community of its contributors, reviewers, and users (Moreau et al., 2009; Open Provenance Model, 2010). The OPM model aims to illustrate the causal relationship between events that impact objects (digital or otherwise) and to elucidate this relationship through a directed acyclic graph (Moreau et al., 2009; Open Provenance Model, 2010). Consequently, researchers studying OPM, in collaboration with the W3C provenance working group, have advanced their research to a new model called PROV (Moreau et al., 2011). According to Groth and Moreau (2013), the PROV document family delineates a model, serializations, and other essential supporting definitions that facilitate the exchange of provenance information in heterogeneous environments, such as the Web. The PROV family of documents comprises four recommendations: the PROV Data Model (PROV-DM), the PROV Ontology (PROV-O), the Provenance Notation (PROV-N), and Constraints of the PROV Data Model (PROV-CONSTRAINTS) (Gil & Miles, 2013; Moreau & Groth, 2013).

#### *1.1.3.2 Data provenance in a general health context and in HIS*

The application of data origin is evident in a wide range of health scenarios, which present challenges in data treatment structures. A notable example is the study by Alvarez et al. (2006), where the application of provenance occurred in the context of organ transplant administration and distribution. The work describes the development of a service-oriented architecture using provenance in medical systems to assist in the decision-making process of an organ transplant. As delineated in Li et al. (2008), an additional initiative that functions in conjunction with health data sources is the Center for Pulmonary Immunity Modeling. This initiative was established through a collaborative effort between the University of Pittsburgh, Carnegie Mellon University, and the University of Michigan. This project entailed the conceptualization and development of a data distribution platform, DataXS, which facilitates the dissemination of experimental data, analyses, and models to participating projects. This project utilizes provenance to maintain a record of the data's provenance, rather than the methodology by which the data were processed. In a recent study, Werder et al. (2022) reported concerns about the provenance of data related to applications of AI recommendations in healthcare. In their study, the authors describe several notable examples, including the use of provenance techniques integrated into EHR systems to predict sepsis, a potentially life-threatening condition in which the body's response to an infection can result in damage to its own tissues. They also discuss the application of data auditing, a practice that can be facilitated by data provenance. This allows healthcare organizations to evaluate the data used to train AI systems and identify potential diseases. Furthermore, they explore the potential of data provenance in health services to enhance understanding of the crucial factors that influence the output of a trained algorithm, such as recommending a specific diagnosis or treatment to the relevant parties.

It is imperative to underscore that, within the health context—particularly in HIS, the tracking of health data provenance empowers patients to maintain complete autonomy over the utilization of their secondary personal data. In essence, this initiative fosters transparency by providing patients with information regarding the utilization of their data in various contexts, including public health surveys, clinical trials, and other health-related

initiatives (Margheri et al., 2020). Current health systems utilize intricate mechanisms to manage provenance, implementing security measures to ensure the authenticity of data sources. However, these approaches are not without their limitations. They are dependent on trusted third parties and are vulnerable to semantic interoperability challenges arising from heterogeneous records maintained by different organizations (Margheri et al., 2020). However, it is imperative to underscore that a multitude of methods, models, and methodologies of data provenance are associated with a diverse array of computational technologies, as delineated in extant literature, to address the particular technological imperatives of HIS. In this context, the application of data provenance—independent of the HIS—provides a fundamental framework for data assessment and verification, thereby ensuring reliability and reproducibility.

#### *1.1.3.3 Data provenance contributing to interoperability in HIS: HL7 FHIR based on W3C PROV*

In the context of data provenance in HIS, it is imperative to underscore that interoperability stands as a pivotal factor to be observed for the optimal functioning of these systems, as it remains a significant challenge that persists. In this sense, HL7 FHIR utilizes provenance as a resource, indicating clinical significance in terms of confidence in the authenticity, reliability, completeness, and lifecycle stage of health data (HL7 International, n.d.). Consequently, HL7 FHIR is predicated on the W3C PROV specification, which delineates mappings of data provenance features. The W3C PROV provides design and implementation means to share semantically interoperable provenance attributes. Moreover, prominent health organizations such as IHE and HL7 endorse the W3C PROV (Margheri et al., 2020). The W3C PROV has been established as the prevailing standard for the representation of interoperable provenance information, having been adopted by HL7 FHIR (Kohlbacher et al., 2018).

#### *1.1.3.4 Main applications of data provenance for the context of HIS*

It is important to note that the concept of data provenance can be applied to a variety of scenarios, including those in the field of health (Cameron, 2003; Pearson, 2002; Sembay et al., 2021). A

considerable body of research in the domain of data provenance has given rise to the development of a taxonomy for the categorization of these efforts, as outlined by Simmhan et al. (2005). As demonstrated in the work of Simmhan et al. (2005), provenance systems can be constructed to function in various ways, exhibiting distinct characteristics and operations. Consequently, this study operates under the assumption that a component of the taxonomy delineated by Simmhan et al. (2005) is indispensable for the examination of the relationships under consideration herein. Data provenance has been shown to have a substantial impact on applications within the context of HIS, as summarized by Goble (2002):(1) Data quality: lineage can be employed to assess data quality and reliability based on the original data and its transformations (Jagadish & Olken, 2004). Additionally, it can serve as proof of data derivation (Silva et al., 2003). (2) Audit trail: provenance enables the tracking of audit trails, determining data usage, and detecting errors in data generation (Galhardas et al., 2001; Greenwood et al., 2003; Miles et al., 2005). (3) Replication recipes: detailed provenance information facilitates the replication of data derivation processes, helps maintain data currency, and acts as a guide for reproduction (Foster et al., 2003; Miles et al., 2005). (4) Attribution: provenance or pedigree can establish copyright and data ownership, enable proper citation, and assign responsibility in cases of erroneous data (Jagadish & Olken, 2004). (5) Informational: a common use of lineage metadata is to support data discovery through queries and browsing, providing contextual information necessary for data interpretation. It is important to emphasize that a deeper understanding of data provenance applications, combined with other emerging technologies, is essential to uncover new opportunities and fully exploit their potential.

### 1.1.4 Blockchain

Blockchain is fundamentally a distributed data structure, frequently referred to as a "public ledger," in which all confirmed transactions are stored in data units known as blocks. Each block in the blockchain contains a reference to the previous block, arranged in chronological order. This arrangement creates a continuous chain that constitutes the blockchain. This chain grows progressively as new transactions are appended to the ledger.

To guarantee the integrity and immutability of the data, blockchain employs asymmetric cryptography, which prevents the alteration of previously recorded blocks (Tian, 2016). Blockchain is an emerging technology that has caused a paradigm shift in various fields on a global scale. The concept was introduced in 2008 with the publication of the white paper "Bitcoin: A Peer-to-Peer Electronic Cash System," which popularized the concept alongside the creation of the Bitcoin cryptocurrency (Nakamoto, 2008). Despite its growing adoption, blockchain remains a complex concept, with multiple definitions emphasizing different aspects of the technology. Swan (2015) categorizes the evolution of blockchain into three distinct phases: (1) Blockchain 1.0: focused primarily on cryptocurrency applications, such as Bitcoin; (2) Blockchain 2.0: expanded applications beyond simple currency transactions to include various types of contracts, such as those related to stocks, loans, mortgages, securities, and smart contracts; (3) Blockchain 3.0: encompasses broader applications extending into domains such as government, healthcare, science, literature, culture, and the arts.

From a technical perspective, blockchain technology facilitates the establishment of a shared, secure, and immutable digital record that chronicles the history of transactions among nodes within public or private peer-to-peer networks. In the context of a transaction, it is imperative that a consensus among all network nodes is achieved to validate and record the transaction. The fundamental purpose of blockchain technology is to establish a decentralized accounting mechanism for transactions, thereby enabling the registration, verification, and transfer of various contracts and assets without the necessity of intermediaries or centralized authorities (Swan, 2015). Beyond its initial implementation in cryptocurrencies, the decentralized and tamper-resistant characteristics of blockchain have facilitated the development of transformative applications in domains such as supply chain management, voting systems, identity verification, and secure medical record-keeping. The potential of blockchain to enhance transparency, security, and trust has led to its recognition as a foundational technology for the future digital economy.

#### *1.1.4.1 Blockchain in a general health context and in HIS*

The potential of blockchain technologies to provide a unique solution for health care is significant. The broad applicability of this technology signifies its potential for integration into diverse facets of medical devices, thereby fostering advancements in various domains of health care. The healthcare sector has seen a mounting demand for blockchain technologies, with established industry players actively exploring novel applications of blockchain to address critical needs (Deloitte, 2018). One of the hallmarks of blockchain, known as immutability, is particularly vital for the storage of health data. This technology has the capacity to safeguard health records and clinical trial results, thereby ensuring regulatory compliance. The utilization of smart contracts exemplifies the application of blockchain technology in facilitating real-time patient monitoring and medical interventions (Griggs et al., 2018). In the domain of health care, blockchain technology exhibits considerable promise in its capacity to disrupt the prevailing methodologies for the management and dissemination of information. This paradigm shift has the potential to profoundly transform existing processes, including the updating and maintenance of medical data, the sharing and synchronization of patient medical records, the assembly and analysis of population health data, and the tracking of prescribed medications throughout the supply chain (Leeming, 2019). Specifically, blockchain has the potential to control access to and distribution of sensitive health information, enhance transparency and auditability of healthcare service delivery, and improve data interoperability across different systems and organizations (Monteil, 2019).

A multitude of studies, including those by Bell et al. (2018) and Zhang et al. (2017), have delineated the pivotal prospective contributions of blockchain technology to the field of health. These objectives include the assurance of data security during health information exchange, the facilitation of nationwide interoperability of health data, and the provision of reliable tracking of medical devices and supply chains. The technology under discussion has been demonstrated to facilitate the monitoring of drug prescriptions, support the surveillance of aggregated health events (leveraging Big Data analytics), and aid in patient identification and secure data sharing for scientific research purposes. Moreover, blockchain technology has the potential to facilitate

the establishment of autonomous and transparent governance structures, such as those necessary for the management and regulation of supplementary health insurance (Bell et al., 2018; Zhang et al., 2017). In the context of HIS, blockchain technology offers innovative solutions that have the potential to enhance the functionality and security of these systems to a considerable degree. At present, EMRS are generally stored in centralized data centers, with access frequently restricted to hospital networks and healthcare providers. This restriction can limit interoperability and patient control over data (Gropper, 2016). Blockchain technology is a decentralized digital ledger that utilizes cryptography for secure and transparent data storage, facilitating comprehensive and tamper-proof patient medical history records.

This approach ensures the immutability and confidentiality of medical records while concurrently enhancing the efficiency of administrative processes. For instance, blockchain has the potential to reduce the time required to resolve insurance claims and improve efficiency in generating insurance quotes by providing transparent and verifiable transaction records. Furthermore, the secure maintenance of patients' comprehensive medical histories through blockchain technology has been demonstrated to facilitate more precise and timely medication recommendations by physicians, thereby enhancing personalized healthcare services and patient safety (Gropper, 2016; Samuel, 2016). The potential applications of blockchain technology in HIS are manifold. These applications include the validation of patient data, the management of EHRS, and the tracking of research methods to manufacture safer medicines. Ensuring proper interoperability, integrity, and privacy of patient information is paramount in all of these applications. Moreover, the implementation of blockchain technology is intended to ensure transparency and auditability in the management of patient information. Most importantly, it seeks to establish robust governance frameworks that ensure proper control, accountability, and secure handling of sensitive health data throughout its lifecycle (Engelhardt, 2017; Kho, 2018; Randall et al., 2017). These mechanisms are critical to fostering trust among patients, healthcare providers, and regulatory bodies, while facilitating compliance with legal and ethical standards.

#### *1.1.4.2 Blockchain contributing to interoperability in HIS*

To maintain patient privacy in the context of data exchange with other institutions within the health ecosystem, it is imperative to implement robust access control mechanisms, ensure the integrity of data provenance, and ensure data interoperability. The interoperability of medical data between healthcare institutions and patient portals on HIS is a promising application of blockchain technology. As this technology matures, its potential to revolutionize all aspects of health care increases, and this is becoming increasingly evident (Hasselgren et al., 2020). A plethora of challenges pertaining to the interoperability of medical data across disparate HIS have been documented in the extant literature. In certain cases, there is a necessity to interweave disparate computational technologies to facilitate the exchange of data between different HIS, either due to institutional policies or the absence of a structured framework in existing standards. Therefore, it is imperative to underscore that the predominant blockchain technology solutions for the interoperability challenges encountered in disparate HIS were examined in numerous articles within the study by Peterson et al. (2016) and Zhang et al. (2018). In these articles, the authors elucidate the interoperability achieved in HIS through the utilization of HL7 FHIR-related features. An alternative approach that was identified involved the implementation of a translator component as a gateway to the data blocks, employing a different standard for translating formats (Roehrs et al., 2017).

#### *1.1.4.3 Blockchain main features*

It is important to highlight that blockchain technology possesses four fundamental features, as outlined by Sultan et al. (2018): (1) Immutable: blockchain acts as a permanent and tamper-proof ledger of transactions. Once a block is added to the chain, it cannot be altered or deleted, thereby ensuring a reliable and verifiable transaction record. (2) Decentralized: the blockchain is stored as a distributed ledger accessible and replicated across multiple nodes in the network. This decentralized architecture eliminates reliance on a central authority, enhancing resilience and reducing single points of failure. (3) Consensus-driven: each block in the blockchain is independently verified through

consensus mechanisms that define specific rules for block validation. These mechanisms often require participants to demonstrate a resource-intensive proof of work or similar effort (as exemplified by Bitcoin mining) to confirm transactions, thereby ensuring trustworthiness without the need for intermediaries. (4) Transparent: blockchain maintains a fully transparent transaction history, which is accessible to all participants in the network. This openness facilitates auditing and creates a provenance trail that allows for comprehensive tracking of the lifecycle and ownership of assets.

### 1.1.5 Related works combining data provenance and blockchain in HIS applications

This section presents related works that combine data provenance and blockchain technology in HIS applications. The selection of studies was made with a focus on two criteria: relevance and alignment with the overarching theme of this research.

1. The initial study, entitled "Integrating blockchain for data sharing and collaboration in mobile healthcare applications" (Liang et al., 2017), is highlighted. The authors proposed an innovative, user-centric solution for health data sharing that leverages a mobile, user-controlled blockchain framework for cloud-based PHR sharing. Their approach employs algorithm-driven techniques for data provenance collection, utilizing blockchain technology. Consequently, the solution incorporates an algorithm capable of managing the provenance of mobile health (mHealth) data while ensuring data integrity and preserving user privacy.

2. The second study is entitled "Using PROV and blockchain to achieve health data provenance" (Massi et al., 2018). The authors propose a decentralized approach to managing healthcare data in EHR systems, grounded in blockchain technologies and the W3C PROV model, as a solution to the prevailing challenges. The solution employed by the aforementioned entities utilizes recognized international standards to ensure the interoperability of health data systems. The proposed framework integrates open systems with blockchain

and the W3C PROV model, thereby enhancing the security, traceability, and immutability of health records.

3. In the third study, titled "Research on personal health data provenance and right confirmation with smart contract," the authors proposed a data provenance model called PROV-Chain. This model was developed to address issues such as data leakage, misuse, and the unauthorized acquisition of personal health information. The PROV-Chain model is built upon blockchain technologies and the OPM (Gong et al., 2019). The model has been designed for PHR applications within the context of IoT environments, with the objective of ensuring secure data sharing and accountability. The evaluation of PROV-Chain demonstrated its effectiveness in ensuring the traceability of personal health data, while also reinforcing users' rights over their own data and enhancing the overall security and integrity of HIS.

4. The fourth study, titled "Secure and provenance enhanced internet of health things framework: A blockchain managed federated learning approach," is situated within the context of the Internet of Health Things (IoHT) (Rayhman et al., 2020), where ensuring data accuracy, security, integrity, and quality is fundamental for stakeholder trust and the effective adoption of IoHT-based solutions. In response to these demands, the authors propose a hybrid federated learning model in which intelligent blockchain-based smart contracts coordinate and manage the training processes. To guarantee complete privacy and anonymity of sensitive IoHT data, the proposed model was evaluated using several machine learning applications developed for clinical trials involving patients with COVID-19. The findings of the study demonstrated that the model effectively preserves data confidentiality while maintaining performance, thereby demonstrating significant potential for the broader adoption of IoHT-based PHR systems in health management.

5. The fifth study, entitled "Decentralised provenance for healthcare data," presents a platform for managing the provenance of EHRs. This platform can be implemented in existing EHR systems (Margheri et al., 2020). The authors

utilize blockchain technology in conjunction with FHIR to represent EHRs. A proxy component transparently intercepts modifications made to EHR and subsequently triggers a smart contract responsible for generating provenance annotations using the W3C PROV standard. These annotations, meticulously structured as PROV documents, are then securely recorded and stored on a hyperledger fabric blockchain. This approach ensures tamper-resistant provenance tracking, thereby enabling transparency, traceability, and verifiability of all changes applied to health records within a decentralized and auditable environment.

Consequently, the related works presented herein will serve as a foundation for some of the analyses carried out in this article, as they are potential studies with the theme addressed here.

## 2 METHODOLOGY

As this article constitutes an expansion of the study by Sembay et al. (2022), the methodology applied herein follows the same premises described, with certain modifications in relation to the incorporation of new analyses for novel reflections. In terms of its nature, this study is classified as basic research, as it is not primarily concerned with immediate application, but rather is embedded in an academic and disciplinary context that focuses on theoretical understanding and analytical rigor (Schauz, 2014).

With regard to the methodological procedures, the research is identified as a bibliographic study, understood as any investigation that involves the collection and analysis of information derived from previously published materials (Allen, 2017). In terms of its objectives, this study assumes an exploratory character, a quality often associated with pilot or feasibility studies. Such studies are essential in assessing the viability and potential value of progressing with a research design or intervention (Hallingberg et al., 2018). Furthermore, the study employs a qualitative approach, which aims to comprehend the dimensions of social reality through nonnumerical data, typically producing and analyzing textual information (McCusker & Gunaydin, 2015). It is also imperative to emphasize that certain analytical and interpretative methodologies employed in this research are rooted

in the frameworks and methodologies developed by Coimbra and Dias (2021) and Gontijo et al. (2021), which serve as the foundation for the critical examination of the selected literature.

To analyze the primary data provenance application relations as defined by Simmhan et al. (2005) with the blockchain features as presented by Sultan et al. (2018), the following features have been considered: (1) Highly relevant: which has a direct effect on data—represented by the symbol ⊠, (2) Relevant: which has an indirect effect on data—represented by the symbol ⊠, and (3) Unidentified: no relation defined—represented by the symbol ⊠. A literature review was conducted to examine the existing connections between data provenance and blockchain, as initially outlined in the foundational works of Simmhan et al. (2005) and Sultan et al. (2018). A comprehensive review of the extant literature was conducted, identifying five studies published between 2017 and 2020 that contributed significantly to the thematic core of this article. The selection of these related works was guided by their alignment—either direct or indirect—with the conceptual relationships proposed in the studies of Simmhan et al. (2005) and Sultan et al. (2018). Accordingly, the analytical framework of this study was structured around the following guiding research questions: (1) What are the existing relationships between data provenance and blockchain technologies? (2) How can the integration of data provenance and blockchain contribute to applications in HIS? (3) What types of data interoperability patterns emerge from the combined use of data provenance and blockchain in HIS? These inquiries were addressed through a meticulous content analysis of the selected literature, thereby facilitating a critical evaluation of the theoretical and practical intersections between provenance, blockchain, and HIS. Consequently, the subsequent section will present analyses that demonstrate how the integration of data provenance and blockchain technology contributes to the success of HIS applications, drawing upon the extant literature on the subject.

## 3 RESULTS

The results of the analyses presented in this section extend the findings originally reported by Sembay et al. (2022).

### *3.1 Identifying the relationships between key applications of data provenance and core blockchain features*

As demonstrated in Table 1, a comparison is presented between the key applications of data provenance and the core features of blockchain technology. This comparison is supported by the findings of studies by Simmhan et al. (2005) and Sultan et al. (2018). The objective of this comparison is twofold: first, to verify the technological compatibility between the two approaches and, second, to identify potential points of convergence. The relationships that were identified are outlined below.

**Table 1.** Identification of the relations between data provenance and blockchain. **Note**. Sembay et al. (2022).

| | | **Core features of blockchain technology** | | | |
|---|---|---|---|---|---|
| | | Transparent | Consensus-driven | Decentralized | Immutable |
| **Key applications of data provenance** | Informational | ◉ | ◉ | ◉ | ◉ |
| | Attribution | ◉ | ◉ | ● | ● |
| | Replication recipes | ● | ◉ | ● | ● |
| | Audit trail | ● | ● | ● | ○ |
| | Data quality | ◉ | ◉ | ◉ | ◉ |

● Highly relevant: which has a direct effect on data;
◉ Relevant: which has an indirect effect on data;
○ Unidentified: no relation defined.

In the study by Sembay et al. (2022), the authors conducted an analysis in Table 1 to identify relationships between the main applications of data provenance and the characteristics of blockchain. The following observations were made: applications related to the informational identity demonstrated relevant relationships with all characteristics (transparent, consensus-driven, decentralized, and immutable). As demonstrated in Table 1, informational applications exhibit relevant connections with all blockchain features (transparent, consensus-driven, decentralized, and immutable), since data discovery benefits from each of these aspects. Attribution applications are closely related to transparent and consensus-driven processes, as they facilitate the establishment of authorship and ownership through the utilization of a verifiable data history. Additionally, strong links have been identified between decentralized systems and those that are immutable, given the paramount importance of accountability in the replication of data and the potential for errors to occur. In the context of replication recipes, consensus-driven mechanisms assume paramount importance, as trust verification ensures the accurate reproduction of data for new experiments. The transparency, decentralization, and immutability of blockchain technology further enhance this process by maintaining an unalterable and distributed record of transactions. In the context of audit trail applications, the attributes of transparency, consensus-driven processes, and decentralization are of paramount importance, as they ensure the traceability and reliability of data across networks. However, a direct correlation with immutable data has not been established. In the context of data quality applications, the full suite of blockchain features is pertinent, as provenance-based quality assurance depends on transparent, immutable, and decentralized data records.

The analysis indicates that audit trail and replication applications exhibit a strong alignment with blockchain capabilities, underscoring the manner in which data provenance, when integrated with blockchain, can enhance data integrity, security, confidentiality, and reliability across multiple domains. In this sense, as indicated in the analysis conducted in the study by Sembay et al. (2022), it is evident that blockchain technology can be employed to minimize aspects related to data provenance, traceability, and data authority guarantee. Indeed, the integration of data provenance with blockchain technologies has been

demonstrated to enhance data reliability and traceability, thereby providing tamper-proof information regarding the origins and transformations of data.

### *3.2 Relations between data provenance and blockchain applied in HIS*

In this section, the related works presented in this article are analyzed based on the study by Sembay et al. (2022). The objective of this analysis is to determine whether the works consider the relationships found in Table 1, specifically applied to HIS. Consequently, Table 2 presents related studies that explore the combined application of data provenance and blockchain technology within HIS.

**Table 2.** Analysis of related works that combine data provenance and blockchain in HIS. **Note**. Sembay et al. (2022).

| **Authors/ years** | **Technologies and frameworks for data provenance** | **Blockchain-based systems** | **Different forms of HIS** | **Identifying the relationships between key applications of data provenance and core blockchain features** |
|---|---|---|---|---|
| Margheri et al. (2020) | W3C PROV | Smart contract/hyperledger fabric blockchain | EHR | Data provenance (informational, replication recipes, audit trail, and data quality) with blockchain (transparent, consensus-driven, decentralized, and immutable) |

| Authors/ years | Technologies and frameworks for data provenance | Blockchain-based systems | Different forms of HIS | Identifying the relationships between key applications of data provenance and core blockchain features |
|---|---|---|---|---|
| Rayhman et al. (2020) | Algorithms based on data provenance | Smart contract | PHR | Data provenance (informational, audit trail, and data quality) with blockchain (consensus-driven, decentralized, and immutable) |
| Gong et al. (2019) | PROV-Chain based on the OPM | Smart contract | PHR | Data provenance (informational and attribution) with blockchain (transparent, consensus-driven, decentralized, and immutable) |
| Massi et al. (2018) | W3C PROV | Blockchain decentralized | EHR | Data provenance (informational, replication recipes, audit trail, and data quality) with blockchain (consensus-driven, decentralized, and immutable) |
| Liang et al. (2017) | Algorithms based on data provenance | Data sharing based on blockchain | PHR | Data provenance (audit trail and data quality) with blockchain (transparent, decentralized, and immutable) |

As demonstrated in Table 2 of the study by Sembay et al. (2022), the analysis emphasizes that data provenance technologies are classified into models—specifically, W3C PROV and OPM—and algorithmic techniques based on data provenance applied to HIS. The W3C PROV model facilitates interoperable exchange of provenance information across heterogeneous environments, such as networks. Its structural definition encompasses entities, activities, and agents engaged in data production or utilization, establishing four fundamental properties: *wasGeneratedBy*, *wasAssociatedBy*, *wasAttributedTo*, and *used* (Gil & Miles, 2013). In contrast, the OPM endeavors to embody provenance for all entities, irrespective of their material or immaterial nature. It does so by elucidating the causal relationships between events that exert an influence on digital or physical objects through a directed acyclic graph (Moreau et al., 2009; Open Provenance Model, 2010). It is important to acknowledge that the OPM model has since been replaced by the W3C PROV standard. With respect to blockchain technologies, the applications are predominantly driven by smart contracts, followed by hyperledger fabric blockchain, blockchain decentralized, and data sharing based on blockchain. With respect to HIS types, the majority of applications target PHR, followed by EHR. Personal health records are frequently established and overseen by patients themselves, with accessibility occurring via desktop computers, web browsers, or mobile devices, including smartphones or portable storage devices (Liu et al., 2011). Conversely, EHRS comprise extensive, interinstitutional, and longitudinal collections of patient health data, which are essential not only for clinical treatment evaluation but also for more comprehensive health management (Hoerbst & Ammenwerth, 2010).

As indicated by Table 2, the relationships identified between data provenance and blockchain across the five reviewed studies correspond closely to those in Table 1, maintaining the same order of relevance. Moreover, the extant literature suggests a growing trend in the adoption of data provenance in conjunction with blockchain technologies within HIS, which contributes positively to health data management. However, it should be emphasized that the scope of the analyzed studies is limited to specific combinations of data provenance and blockchain technologies applied to HIS, as aligned with the article's focus. It is noteworthy that other literature may present alternative combined technologies and successful implementations in various health contexts. This

assertion is supported by studies such as Puel et al. (2014), Macedo et al. (2015, 2019), and Sembay et al. (2023). Consequently, as illustrated in Table 2 of the Sembay et al. (2022) study, it is imperative to acknowledge the potential of a combined approach involving data provenance and blockchain in HIS. This integration has the capacity to induce alterations within the ecosystem of the health sector, thereby fostering trust and enhancing efficiency, thus leading to an improvement in patient treatment. Additionally, it facilitates the secure and transparent dissemination of health information stored in the HIS, thereby enhancing the accessibility of these data to external health institutions that require them to continue patient treatment. In this manner, the blockchain provides the requisite resources to guarantee data provenance in HIS. Nevertheless, challenges may arise due to technological factors, yet these present more advantages than disadvantages.

### *3.3 Main standards of interoperability found between data provenance and blockchain*

In this section, the related works previously described are analyzed with respect to the use of the primary data interoperability standards in HIS. Consequently, Table 3 provides a comprehensive analysis of the study's findings.

**Table 3.** Analysis of the main standards of interoperability found in related works that combine data provenance and blockchain in HIS. **Note**. Prepared by the authors.

| Authors/Years | Types of HIS | Main standards of interoperability used |
|---|---|---|
| Margheri et al. (2020) | EHR | HL7 FHIR, IHE, DICOM, XDS |
| Rayhman et al. (2020) | PHR | ETL |
| Gong et al. (2019) | PHR | ETL |
| Massi et al. (2018) | EHR | HL7 FHIR, IHE, DICOM, XDS, CDA |

| Authors/Years | Types of HIS | Main standards of interoperability used |
|---|---|---|
| Liang et al. (2017) | PHR | ETL |

In the study by Margheri et al. (2020), the authors present the importance of utilizing HL7 FHIR, IHE, DICOM, and XDS in EHR. The authors posit that these interoperability standards facilitate the formulation of strategies by policymakers and project coordinators, ensuring software sustainability and safeguarding investments, while concurrently enhancing patient data security and the quality of care provided by healthcare institutions. Additionally, the use of these standards is said to optimize the combined use of data provenance and blockchain technologies in the health services offered by the HIS. In the study by Rayhman et al. (2020), the authors report the use of mHealth devices in PHR in the context of the IoHT, based on the ETL standard. This contributes to the collection of health data from various sources of mobile devices, its transformation according to the needs of the database, and its loading into a database where the necessary correlations occur for the use of these health data by the specialist professional. In this regard, the utilization of the ETL standard in the PHR facilitates the consolidation and presentation of transaction data from a data warehouse or other health database, ensuring its perpetual availability for viewing. Consequently, this enables the efficacy of data provenance processes in conjunction with blockchain technologies, particularly in IoHT scenarios, within the context of HIS.

In the study by Gong et al. (2019), the authors explore the integration of mHealth devices within hospital settings, particularly within the context of PHRs. The utilization of ETL processes plays a pivotal role in this context, facilitating the upload of comprehensive health data from these devices. Additionally, ETL contributes to the extraction of data, the maintenance of a copy of the most recent extraction, and the subsequent transfer of these data to a secure health database. Consequently, the utilization of ETL for the processes involved in data provenance and blockchain applied in the context of PHR contributes to the tracking of health data in these scenarios. In the study by Massi et al. (2018), the authors mention the use of HL7 FHIR, IHE, DICOM, XDS, and

CDA to contribute to the existing limitations of interoperability in EHR systems. The authors posit that the established criteria contribute to the system they have proposed, which utilizes blockchain technology to manage the provenance of health documents. This system is designed to seamlessly integrate into existing EHR deployments. In the study conducted by Liang et al. (2017), the authors posited that the dissemination of health data among institutions necessitates the establishment of a secure data sharing infrastructure. However, there are several challenges related to privacy, security, and interoperability. In this regard, the utilization of the ETL standard for mHealth devices to populate health databases plays a pivotal role in facilitating the implementation of blockchain technologies for the management of data provenance in PHR and the development of novel iterations of EHR, featuring user-centric access control and privacy preservation mechanisms.

Therefore, the interoperability standards highlighted in each study analyzed in Table 3 demonstrate that they are critical requirements for HIS. Notwithstanding the extant limitations that may imperil patient safety, the interoperability standards delineated in Table 3 are the most widely utilized and contribute to ameliorating the preponderance of limitations in the exchange of health data between disparate HIS. In conclusion, an evident correlation was identified among the patterns exhibited in Table 3. These patterns collectively contributed to enhancing the requirements concerning the tracking of health data, as well as the security and immutability of these data when utilizing blockchain technologies in HIS.

### *3.4 Summary and reflections of the analysis presented*

This section presents a synopsis of the analysis performed, as illustrated in Figure 1. The analysis was conducted using data from Tables 1–3, and it highlights the significance of the elements identified during the course of the study.

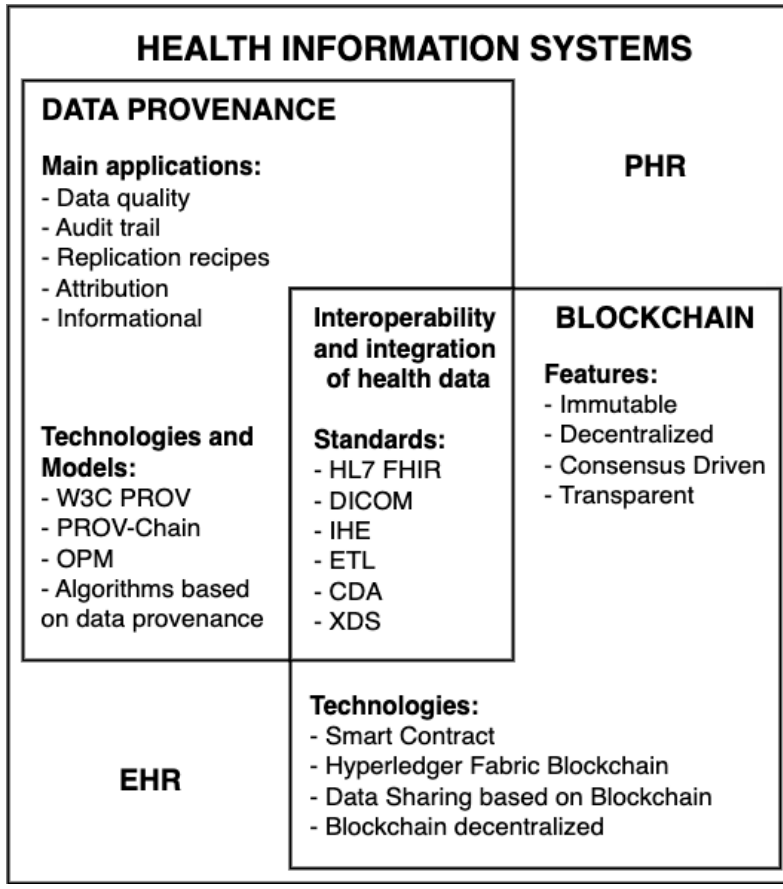


**Figure 1.** Summary of analysis. **Note**. Prepared by the authors.

Figure 1 presents the primary elements identified in the analysis, underscoring the significance of integrating data provenance and blockchain within the framework of HIS. With regard to Figure 1, the following observations can be made: (1) It was observed in the studies that the EHR and PHR are, in fact, the most used HIS. Making a general analysis of these two systems, the following reflections stand out: the EHR is the most used by doctors to improve the quality of care, having as its main advantage the

availability of medical information between providers; the EHR and PHR reside on different platforms under various technologies and standards; and PHR allows the integration of the main information components in the EHR systems. Thus, it is important to emphasize that the integration of medical information into EHR and PHR leads to a dramatic change in personalized care; (2) Regarding the main applications of data provenance (data quality, audit trial, replication recipes, attribution, and informational) that intertwine with blockchain characteristics (immutable, decentralized, consensus-driven, and transparent) in the context of HIS, it can be stated that data provenance and blockchain when combined in the context of HIS, mainly in EHR and PHR, result in benefits for this context. In this sense, data provenance is the foundation of medical data quality and patient privacy, and blockchain contributes to the creation and management of provenance records, both in the context of EHR and PHR; and (3) Regarding data provenance technologies and models (W3C PROV, PROV-Chain, OPM, and algorithms based on data provenance) together with blockchain technologies (smart contract, hyperledger fabric blockchain, data sharing based on blockchain, and blockchain decentralized) result in several challenges encountered in data interoperability issues in EHR and PHR. A significant challenge confronting the field is the establishment of system interoperability, defined as the standardization of data and information that can be read, understood, and accessed from any health unit, whether public or private. In this sense, the utilization of these technologies and models of data provenance and blockchain underscores the nexus that pertains to interoperability and integration of health data (HL7 FHIR, DICOM, IHE, ETL, CDA, and XDS).

While these standards do not address all interoperability issues, they aim to align with the requirements of EHR and PHR, thereby enhancing the quality of care and the efficiency of healthcare services. Furthermore, these standards are designed to facilitate the secure exchange of health data between EHRS and PHRS, thereby promoting interoperability and enhancing patient care.

## 4 DISCUSSION

The findings of this study underscore the significance of integrating data provenance and blockchain technology to enhance efficiency, security, and interoperability in HIS, particularly within the domains of EHR and PHR. As the analysis indicates, EHRS remain the primary instrument utilized by healthcare professionals to ensure the delivery of quality care. Conversely, PHRS promote the integration of patient information across multiple platforms, thereby fostering a more personalized care approach. The intersection of data provenance and blockchain has demonstrated considerable potential. Provenance contributes attributes such as traceability, auditing, information quality, and attribution of authorship, which are essential for guaranteeing the integrity of medical data. The blockchain technology under discussion in this paper is characterized by its immutability, decentralization, and transparency. These characteristics contribute to the establishment of a robust layer of security and reliability. The integration of these technologies, as Sembay et al. (2022) have noted, has been demonstrated to enhance the creation of reliable and auditable records, thereby reducing risks and strengthening confidence in the use of HIS. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The analysis identified that, despite the potential of these technologies, significant challenges related to interoperability persist. The heterogeneity of standards and platforms adopted by EHR and PHR systems engenders challenges in the seamless integration of data. Technologies such as W3C PROV, PROV-Chain, OPM, smart contracts, and hyperledger fabric blockchain, when associated with interoperability standards such as HL7 FHIR, DICOM, IHE, CDA, and XDS, seek to mitigate these challenges. Despite these advancements, the pursuit of complete standardization remains unfinished, as technical and institutional barriers persist. From a pragmatic standpoint, the findings of this study underscore the feasibility of enhancing health data integration through the joint implementation of data provenance and blockchain technology. This approach is proposed as a method to guarantee the integrity, security, and effective dissemination of information. The study makes a theoretical contribution

to the field by offering insights into the complementarity between these technologies. This understanding can inform the development of future models for more secure and interoperable HIS architecture. From a political standpoint, the findings underscore the necessity for public and regulatory policies that promote the utilization of open standards and reliable technologies for health data management.

However, it is important to note that this research is not without its limitations. The analysis was primarily based on secondary studies and a limited sample of related work. It should be noted that no empirical experiments or direct surveys were conducted in actual HIS environments. Furthermore, given the perpetual evolution of technology, it is important to note that the results may not fully capture the most recent advancements. These factors may limit the generalizability of the findings and necessitate caution when extrapolating the results. In summary, the findings suggest that integrating data provenance with blockchain technology holds potential for enhancing the quality, security, and interoperability of HIS. The future of this field will be determined by three factors: greater standardization, practical experimentation, and collaboration between technological agents, health professionals, and public policy makers.

## 5 CONCLUSION

The analysis described in this article suggests that data provenance applications combined with blockchain have the potential to be promising in a variety of application sectors, as illustrated in Table 1. In this sense, as illustrated in Table 1, it was possible to understand that the relationships found may be directed to the HIS, specifically the EHR and PHR, as shown in Table 2 in the analysis carried out in the works presented. Therefore, as indicated by the findings presented in Table 2, it is evident that alterations in the EHR and PHR ecosystem may transpire. To address this, it is imperative to identify suitable models, methods, techniques, and methodologies that will empower health organizations to store provenance records. These records, in turn, must be shared and tracked by the blockchain structure, thereby mitigating the risk of data tampering. This approach serves to mitigate the complexity encountered by HIS when confronted with

substantial volumes of health data, which necessitates secure and reliable management. The integration of blockchain technology with data provenance holds considerable promise in this regard. Furthermore, an analysis of Table 2 suggests that the most prominent HIS in the studies are: EHR and PHR. This is because the EHR is used as the standard medical record used in several countries in their respective HIS, and the PHR is the most convenient for patients and healthcare professionals who can monitor health data remotely via mobile devices, especially in times of pandemic, as was the case with COVID-19.

Another salient point pertains to the data interoperability standards delineated in Table 3, which concerns the amalgamation of data provenance and blockchain in HIS, particularly in the context of EHR and PHR. These standards contribute to the normalization and interoperability of health data in the aforementioned HIS. However, challenges persist, particularly with regard to the security and privacy of patient data. This indicates that, given the existence of multiple HIS, health institutions have prioritized standardizing clinical procedures to ensure uniformity in practice and establishing systems to facilitate the exchange of data and information across different HIS. The analysis, as depicted in Figure 1, demonstrates that the integration of data provenance and blockchain in EHR and PHR systems, despite the challenges associated with this integration, offers significant benefits. Figure 1 underscores a mounting trend in the implementation of blockchain technology for the management of healthcare document provenance, exhibiting the capacity for seamless integration across disparate healthcare institutions. This approach facilitates the secure management of document provenance while ensuring data privacy.

Finally, as a suggestion for future research, it is recommended to undertake a more comprehensive and detailed systematic literature review that extends beyond the scope of this study. A rigorous investigation should be undertaken to ascertain the existence of any additional applications and integrations of data provenance combined with blockchain technology across various HIS. This includes a thorough presentation of the principal methods, techniques, models, and methodologies employed in conjunction with data provenance and blockchain within different HIS contexts. Furthermore, subsequent research endeavors should investigate the operational dynamics of integrating data

provenance and blockchain within HIS environments that incorporate advanced technologies, such as cloud computing and the IoHT. Additionally, these studies should examine their interactions with AI applications in medicine. It is imperative to comprehend these relationships to ascertain how this technological synergy can enhance data security, interoperability, transparency, and decision-making processes in healthcare. Indeed, a more in-depth study on the combined use of data provenance and blockchain has the potential to offer significant contributions by clarifying the specific challenges faced in these implementations. Furthermore, it would facilitate the identification of the most effective technological architectures and frameworks that maximize the benefits of this integration, thereby supporting the successful deployment and adoption of these innovations across diverse HIS.

## *Conflict of interest*

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## *Contribution statement*

**Márcio José Sembay:** Writing – Review & Editing, Writing – Original Draft.
**Alexandre Augusto Gimenes Marquez Filho:** Methodology, Conceptualization.
**Douglas Dyllon Jeronimo de Macedo:** Writing – Review & Editing, Supervision.

## *Statement of data consent*

This study did not generate any new or large-scale datasets requiring deposition in repositories. The research is based exclusively on the analysis and interpretation of related works previously published in the literature, which supported the development and extension of this manuscript.

## REFERENCES


Al Jarullah, A., & El-Masri, S. (2012). Proposal of an architecture for the national integration of electronic health records: A semi-centralized approach. *Studies in Health Technology and Informatics, 180*, 917–921. *https://doi.org/10.3233/978-1-61499-101-4-917*

Allen, M. (2017). Bibliographic research. In *The SAGE Encyclopedia of communication research methods*. SAGE Publications. *https://doi.org/10.4135/9781483381411.n37*

Alvarez, S., Vazquez-Salceda, J., Kifor, T., Varga, L. Z., & Willmott, S. (2006). Applying provenance in distributed organ transplant management. In *Provenance and annotation of data: International provenance and annotation workshop, IPAW 2006, Chicago, IL, USA, May 3–5, Revised Selected Papers* (pp. 28–36). Springer Berlin Heidelberg. *https://doi.org/10.1007/11890850_4*

Andargolia, A. E., Scheepers, H., Rajendran, D., & Sohal, A. (2017). Health information systems evaluation frameworks: A systematic review. *International Journal of Medical Informatics, 97*, 195–209. *https://doi.org/* 10.1016/j.ijmedinf.2016.10.008

Annas, G. J. (2003). HIPAA regulations: A new era of medical-record privacy? *New England Journal of Medicine, 348*(15), 1486. *https://doi.org/10.1056/NEJMlim035027*

Bakker, A. R. (1991). HIS, RIS, and PACS. *Computerized Medical Imaging and Graphics, 15*(3), 157–160. *https://doi.org/10.1016/0895-6111(91)90004-F*

Bansal, S. K. (2014). Towards a semantic extract-transform-load (ETL) framework for big data integration. In *2014 IEEE international congress on big data* (pp. 522–529). IEEE. http://10.1109/BigData.Congress.2014.82

Bell, L., Buchanan, W. J., Cameron, J., & Lo, O. (2018). Applications of blockchain within healthcare. *Blockchain in Healthcare Today, 1*, 1–7. *https://doi.org/10.30953/bhty.v1.8*

Bernardini, A., Alonzi, M., Campioni, P., Vecchioli, A., & Marano, P. (2003). IHE: Integrating the Healthcare Enterprise, towards complete integration of healthcare information systems. *Rays, 28*(1), 83–93. *https://pubmed.ncbi.nlm.nih.gov/14509182/*

Blick, K. E. (1997). Decision-making laboratory computer systems as essential tools for achievement of total quality. *Clinical Chemistry, 43*(5), 908–912. *https://doi.org/10.1093/clinchem/43.5.908*

Boochever, S. S. (2004). HIS/RIS/PACS integration: Getting to the gold standard. *Radiology Management, 26*(3), 16–24. *https://pubmed.ncbi.nlm.nih.gov/15259683/*

Buneman, P., Khanna, S., & Wang-Chiew, T. (2001). Why and where: A characterization of data provenance. In J. Van den Bussche & V. Vianu (Eds.), *ICDT 2001: International conference on database theory* (pp. 316–330). Springer. *https://doi.org/10.1007/3-540-44503-X_20*

Cameron, G. (2003). *Provenance and pragmatics* [Workshop on Data Provenance and Annotation]. Edinburgh, UK.

Cesnik, B., & Kidd, M. R. (2010). History of health informatics: A global perspective. *Studies in Health Technology and Informatics, 151*, 3–8. *https://doi.org/10.3233/978-1-60750-476-4-3*

Coimbra, F. S., & Dias, T. M. R. (2021). Use of open data to analyze the publication of articles in scientific events. *Iberoamerican Journal of Science Measurement and Communication, 1*(3), 1–13. *https://doi.org/10.47909/ijsmc.123*

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. *Journal of Big Data, 6*(1), 1–25. *https://doi.org/10.1186/s40537-019-0217-0*

Dehnavieh, R., Haghdoost, A., Khosravi, A., Hoseinabadi, F., Rahimi, H., Poursheikhali, A., Shafiee, G., Gholami, H., Abadi, M. B. H., Noori, R., & Mehrolhassani, M. H. (2018). The District Health Information System (DHIS2): A literature review and meta-synthesis of its strengths and operational challenges based on the experiences of 11 countries. *Health Information Management Journal, 48*(2), 62–75. *https://doi.org/10.1177/1833358318777713*

Deloitte. (2018). *Breaking blockchain open: Deloitte's 2018 global blockchain survey* (Report No. 48). Deloitte Insights. *https://doi.org/10.1002/ejoc.201200111*

Dolin, R. H., Alschuler, L., Beebe, C., Biron, P. V., Boyer, S. L., Essin, D., & Mattison, J. E. (2001). The HL7 clinical document architecture. *Journal of the American Medical Informatics Association, 8*(6), 552–569. *https://doi.org/10.1136/jamia.2001.0080552*

Engelhardt, M. A. (2017). Hitching healthcare to the chain: An introduction to blockchain technology in the healthcare sector. *Technology Innovation Management Review, 7*(10), 22–34. *http://doi.org/10.22215/timreview/1111*

Foster, I. T., Vöckler, J.-S., Wilde, M., & Zhao, Y. (2003). *The virtual data grid: A new model and architecture for data-intensive collaboration.* In *15th International conference on scientific and statistical database management (SSDBM), Cambridge, MA, USA*. *https://www.cidrdb.org/cidr2003/program/p18.pdf*

Freire, J., Silva, C. T., Callahan, S. P., Santos, E., Scheidegger, C. E., & Vo, H. T. (2008). Provenance for computational tasks: A survey. *Journal of Computing Science and Engineering, 10*(3), 11–21. *https://doi.org/10.1109/MCSE.2008.79*

Friedman, C., Rubin, J., Brown, J., Buntin, M., Corn, M., Etheredge, L., Gunter, C., Musen, M., Platt, R., Stead, W., Sullivan, K., & Van Houweling, D. (2015). Toward a science of learning systems: A research agenda for the high-functioning learning health system. *Journal of the American Medical Informatics Association, 22*(1), 43–50. *https://doi.org/10.1136/amiajnl-2014-002977*

Galhardas, H., Florescu, D., Shasha, D., Simon, E., & Saita, C. A. (2001). *Improving data cleaning quality using a data lineage facility*. In *Proceedings of the international workshop on design and management of data warehouses (DMDW), Interlaken, Switzerland* (pp. 1–13). *http://ceur-ws.org/Vol-39/paper3.pdf*

Gil, Y., & Miles, S. (2013). *PROV model primer* [W3C Working Draft]. W3C. *https://www.w3.org/TR/prov-primer/*

Goble, C. (2002). *Position statement: Musings on provenance, workflow and (Semantic Web) annotations for bioinformatics* [Workshop on Data Derivation and Provenance]. Chicago, IL, USA.

Gong, J., Lin, S., & Li, J. (2019). Research on personal health data provenance and right confirmation with smart contract. In *IEEE 4th advanced information technology, electronic and automation control conference (IAEAC). https://doi.org/10.1109/IAEAC47372.2019.8997930*

Gontijo, M. C. A., Hamanaka, R. Y., & de Araujo, R. F. (2021). Research data management: A bibliometric and altmetric study based on Dimensions. *Iberoamerican Journal of Science Measurement and Communication, 1*(3), 1–19. *https://doi.org/10.47909/ijsmc.120*

Greenspan, G. (2016). *Four genuine blockchain use cases* [Technical report]. MultiChain. *https://www.multichain.com/blog/2016/05*

Greenwood, M., Goble, C., Stevens, R., Zhao, J., Addis, M., Marvin, D., Moreau, L., & Oinn, T. (2003). *Provenance of e-science experiments: Experience from bioinformatics*. In *Proceedings of the UK OST e-science second all hands meeting, Nottingham, UK. https://eprints.soton.ac.uk/258895/1/prov-all-hands.pdf*

Griggs, K. N., Ossipova, O., Kohlios, C. P., Baccarini, A. N., Howson, E. A., & Hayajneh, T. (2018). Healthcare blockchain system using smart contracts for secure automated remote patient monitoring. *Journal of Medical Systems, 42*(7), 130. *https://doi.org/10.1007/s10916-018-0982-x*

Gropper, A. (2016). *Powering the physician-patient relationship with HIE of one blockchain health IT* [ONC/NIST Use of Blockchain for Healthcare and Research Workshop]. Gaithersburg, MD. *https://www.healthit.gov/sites/default/files/7-29-poweringthephysician-patientrelationshipwithblockchainhealthit.pdf*

Groth, P., & Moreau, L. (2013). *PROV-overview: An overview of the PROV family of documents*. W3C. *https://www.w3.org/TR/prov-overview/*

Hallingberg, B., Turley, R., Segrott, J., Wight, D., Craig, P., Moore, L., Murphy, S., Robling. M., Simpson, S. A., & Moore, G. (2018). Exploratory studies to decide whether and how to proceed with full-scale evaluations of public health interventions: A systematic review of guidance. *Pilot and Feasibility Studies, 4*, 104. *https://doi.org/10.1186/s40814-018-0290-8*

Hasselgren, A., Kralevska, K., Gligoroski, D., Pedersen, S. A., & Faxvaag, A. (2020). Blockchain in healthcare and health sciences: A scoping review. *International Journal of Medical Informatics, 134*, Article 104040. *https://doi.org/10.1016/j.ijmedinf.2019.104040*

Haux, R. (2006). Health information systems—Past, present, future. *International Journal of Medical Informatics*, *75*(3–4), 268–281. *https://doi.org/10.1016/j.ijmedinf.2005.08.002*

HL7 International. (n.d.). *Fast healthcare interoperability resources release (STU)*. *http://fhir.hl7.org*

Hoerbst, A., & Ammenwerth, E. (2010). Electronic health records. *Methods of Information in Medicine, 49*(4), 320–336. *https://doi.org/10.3414/me10-01-0038*

Honeyman, J. C. (1999). Information systems integration in radiology. *Journal of Digital Imaging, 12*(Suppl 1), 218–219. http://doi: 10.1007/BF03168810

Huang, H. K. (2019). *PACS-based multimedia imaging informatics: Basic principles and applications* (3rd ed.). John Wiley & Sons. *https://doi.org/10.2345/i0899-8205-40-2-125.1*

Ismail, A., Jamil, A. T., Rahman, A. F. A., Bakar, J. M. A., Saad, N. M., & Saadi, H. (2010). The implementation of Hospital Information System (HIS) in tertiary hospitals in Malaysia: A qualitative study. *Malaysian Journal of Public Health Medicine, 10*(2), 16–24.

Jagadish, H. V., & Olken, F. (2004). Database management for life sciences research. *ACM SIGMOD Record, 33*(2), 15–20. *https://doi.org/10.1145/1024694.1024697*

Kho, W. (2018). Blockchain revolution in healthcare: The era of patient-centred dental information system. *International Journal of Oral Biology, 43*(1), 1–3. *https://doi.org/10.11620/IJOB.2018.43.1.001*

Kohlbacher, O., Mansmann, U., Bauer, B., Kuhn, K., & Prasser, F. (2018). Data Integration for Future Medicine (DIFUTURE): An architectural and methodological overview. *Methods of Information in Medicine, 57*(S01), e43–e50. *https://doi.org/10.3414/ME17-02-0022*

Korhonen, I., Pärkkä, J., & van Gils, M. (2003). Health monitoring in the home of the future. *IEEE Engineering in Medicine and Biology Magazine, 22*(3), 66–73. *https://doi.org/10.1109/MEMB.2003.1213628*

Law, M. Y., & Zhou, Z. (2003). New direction in PACS education and training. *Computerized Medical Imaging and Graphics, 27*(2–3), 147–156. *https://doi.org/10.1016/S0895-6111(02)00088-5*

Leeming, G., Ainsworth, J., & Clifton, D. A. (2019). Blockchain in health care: Hype, trust, and digital health. *The Lancet*, 393(10190), 2476–2477. *https://doi.org/10.1016/S0140-6736(19)30948-1*

Li, Q., Labrinidis, A., & Chrysanthis, P. K. (2008). User-centric annotation management for biological data. In *Provenance and annotation of data and processes: second international provenance and annotation workshop, IPAW 2008, Salt Lake City, UT, USA, June 17–18, Revised Selected Papers* (pp. 54–61). Springer Berlin Heidelberg. *https://doi.org/10.1007/978-3-540-89965-5_7*

Liang, X., Zhao, J., Shetty, S., Liu, J., & Li, D. (2017). *Integrating blockchain for data sharing and collaboration in mobile healthcare applications.* In *2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC), Montreal, QC, Canada* (pp. 1–25). IEEE. *https://doi.org/10.1109/PIMRC.2017.8292361*

Liu, L. S., Shih, P. C., & Hayes, G. (2011). Barriers to the adoption and use of personal health record systems. *Proceedings of the iConference*, 363–370. *https://doi.org/10.1145/1940761.1940811*

Macedo, D. D. J., de Von Wangenheim, A., & de Dantas, M. A. R. (2015). A data storage approach for large-scale distributed medical systems. In *2015 Ninth international conference on complex, intelligent, and software intensive systems* (pp. 486–490). *https://doi.org/10.1109/CISIS.2015.88*

Macedo, D. D., de Araújo, G. M., de Dutra, M. L., Dutra, S. T., & Lezana, Á. G. (2019). Toward an efficient healthcare Cloud IoT architecture by using a game theory approach. *Concurrent Engineering, 27*(3), 189–200. *https://doi.org/10.1177/1063293X19844548*

Margheri, A., Massi, M., Miladi, A., Sassone, V., & Rosenzweig, A. J. (2020). Decentralised provenance for healthcare data. *International Journal of Medical Informatics, 141*, Article 104197. *https://doi.org/10.1016/j.ijmedinf.2020.104197*

Massi, M., Miladi, A., Margheri, A., Sassone, V., & Rosenzweig, J. (2018). *Using PROV and blockchain to achieve health data provenance* [Technical Report]. University of Southampton. *https://eprints.soton.ac.uk/421292/1/PROV_BC_Healthcare.pdf*

McCusker, K., & Gunaydin, S. (2015). Research using qualitative, quantitative or mixed methods and choice based on the research. *Perfusion, 30*(7), 537–542. *https://doi.org/10.1177/0267659114559116*

Mildenberger, P., Eichelberg, M., & Martin, E. (2002). Introduction to the DICOM standard. *European Radiology, 12*(4), 920–927. http://doi: 10.1007/s003300101100

Miles, S., Groth, P., Branco, M., & Moreau, L. (2005). *The requirements of recording and using provenance in eScience experiments* [Technical Report]. Electronics and Computer Science, University of Southampton, UK. *https://eprints.soton.ac.uk/260269/1/pasoa04requirements.pdf*

Monteil, C. (2019). Blockchain and health. In *Digital medicine* (pp. 41–47). Springer. *https://doi.org/10.1007/978-3-319-98216-8_4*

Moreau, L. (2006). *Usage of "provenance": A tower of Babel* [Position paper]. Microsoft Life Cycle Seminar, Mountain View, CA. *https://eprints.soton.ac.uk/409382/*

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., & Van den Bussche, J. (2011). The open provenance model core specification (v1.1). *Future Generation Computer Systems, 27*(6), 743–756. *https://doi.org/10.1016/j.future.2010.07.005*

Moreau, L., & Groth, P. (2013). *Provenance: An introduction to PROV*. Morgan & Claypool. *https://doi.org/10.2200/s00528ed1v01y201308wbe007*

Moreau, L., Kwasnikowska, N., & Van den Bussche, J. (2009). *The foundations of the open provenance model*. University of Southampton. *https://eprints.soton.ac.uk/267282/1/fopm.pdf*

Nadkarni, P. M., Marenco, L., & Brandt, C. (2012). Clinical research information systems. In *Health informatics* (pp. 135–154). Springer. *https://doi.org/10.1007/978-1-84882-448-5_8*

Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system.* National Intelligence Council. *https://fas.org/irp/nic/disruptive.pdf*.

Noumeir, R., & Renaud, B. (2010). IHE cross-enterprise document sharing for imaging: Interoperability testing software. *Source Code for Biology and Medicine, 5*(1), 1–15. *https://doi.org/10.1186/1751-0473-5-9*

Oosterwijk, H. (2002). *DICOM basics* (2nd ed.). Otech.

Open Provenance Model (OPM). (2010). *Open Provenance Model (OPM) specifications*. *https://openprovenance.org/opm/old-index.html*

Pearson, D. (2002). *Presentation on grid data requirements scoping metadata & provenance* [Workshop on Data Derivation and Provenance], Chicago, IL, USA.

Peterson, K., Deeduvanu, R., Kanjamala, P., & Boles, K. (2016). *A blockchain-based approach to health information exchange networks.* U.S. Department of Health and Human Services. *https://www.healthit.gov/sites/default/files/12-55-blockchain-based-approach-final.pdf*

Puel, A., Wangenheim, A. V., Meurer, M. I., & de Macedo, D. D. J. (2014). BUCOMAX: Collaborative multimedia platform for real-time manipulation and visualization of bucomaxillofacial diagnostic images. In *2014 IEEE 27th international symposium on computer-based medical systems* (pp. 392–395). *https://doi.org/10.1109/CBMS.2014.12*

Randall, D., Goel, P., & Abujamra, R. (2017). Blockchain applications and use cases in health information technology. *Journal of Health & Medical Informatics, 8*(3), 1–17. *https://doi.org/10.4172/2157-7420.1000276*

Rayhman, M. A., Hossain, M. S., Islam, M. S., Alrajeh, N. A., & Muhammad, G. (2020). Secure and provenance enhanced internet of health things framework: A blockchain managed federated learning approach. *IEEE Access, 8*, 205071–205087. *https://doi.org/10.1109/ACCESS.2020.3037474*

Robertson, A., Cresswell, K., Takian, A., Petrakaki, D., Crowe, S., Cornford, T., Barber, N., Avery, A., Fernando, B., Jacklin, A., Prescott, R., Klecun, E., Paton, J., Lichtner, V., Quinn, C., Ali, M., Morrison, Z., Jani, Y., Waring, J., Marsden, K., & Sheikh, A. (2010). Implementation and adoption of nationwide electronic health records in secondary care in England: Qualitative analysis of interim results from a prospective national evaluation. *BMJ, 341*, Article c4564. *https://doi.org/10.1136/bmj.c4564*

Roehrs, A., da Costa, C. A., & da Righi, R. R. (2017). Omni**PHR**: A distributed architecture model to integrate personal health records. *Journal of Biomedical Informatics, 71*, 70–81. *https://doi.org/10.1016/j.jbi.2017.05.012*

Samuel, A. M., & Garcia-Constantino, M. (2022). User-centred prototype to support wellbeing and isolation of software developers using smartwatches. *Advances in Notes in Information Science, 1*, 140–151. *https://doi.org/10.47909/anis.978-9916-9760-0-5.125*

Samuel, R. E. (2016). A layered architectural approach to understanding distributed cryptographic ledgers. *Issues in Information Systems, 17*(IV), 222–226. *https://doi.org/10.48009/4_iis_2016_222-226*

Schauz, D. (2014). What is basic research? Insights from historical semantics. *Minerva, 52*(3), 273–328. *https://doi.org/10.1007/s11024-014-9255-0*

Sembay, M. J., de Macedo, D. D. J., & Dutra, M. L. (2021). A proposed approach for provenance data gathering. *Mobile Networks and Applications, 26*(1), 304–318. *https://doi.org/10.1007/s11036-020-01648-7*

Sembay, M. J., de Macedo, D. D. J., Júnior, L. P., Braga, R. M. M., & Sarasa-Cabezuelo, A. (2023). Provenance data management in health information systems: A systematic literature review. *Journal of Personalized Medicine, 13*(6), 991. *https://doi.org/10.3390/jpm13060991*

Sembay, M. J., de Macedo, D. D. J., & Marquez Filho, A. A. G. (2022). Identification of the relationships between data provenance and blockchain as a contributing factor for health information systems. In *Proceedings of data and information in online environments: third eai international conference, DIONE 2022* (pp. 258–272). Springer Nature Switzerland. *http://doi.org/10.1007/978-3-031-22324-2_20*

Sembay, M. J., & Macedo, D. D. J. (2022). Health information systems: proposal of a provenance data management methodinthe instantiation of the W3C PROV-DM model. *Advances in Notes in Information Science, 2*, 101. ColNes Publishing. *https://doi.org/10.47909/anis.978-9916-9760-3-6.101*

Sembay, M. J., Macedo, D. D., & Dutra, M. L. (2020). A method for collecting provenance data: A case study in a Brazilian hemotherapy center. In *Proceedings of the 1st EAI international conference on data and information in online environments (DIONE 2020)* (pp. 1–14). EAI. *https://doi.org/10.1007/978-3-030-50072-6_8*

Silva, P. P. da, Silva, D., McGuinness, D. L., & McCool, R. (2003). Knowledge provenance infrastructure. *IEEE Data Engineering Bulletin, 26*(4), 26–32. *https://dspace.rpi.edu/items/cd532a33-7392-4046-a4a2-c71679ec66eb*

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). *A survey of data provenance techniques* [Technical Report No. TR-618]. Computer Science Department, Indiana University. *https://legacy.cs.indiana.edu/ftp/techreports/TR618.pdf*

Sligo, J., Gauld, R., Roberts, V., & Villac, L. (2017). A literature review for large-scale health information system project planning, implementation and evaluation. *International Journal of Medical Informatics, 97*, 86–97. *https://doi.org/10.1016/j.ijmedinf.2016.09.007*

Sultan, K., Ruhi, U., & Lakhani, R. (2018). *Conceptualizing blockchains: Characteristics & applications*. arXiv. *https://arxiv.org/abs/1806.03693*

Swan, M. (2015). *Blockchain: Blueprint for a new economy*. O'Reilly Media.

Tan, W. C. (2008). Provenance in databases: Past, current, and future. *IEEE Data Engineering Bulletin, 30*(4), 3–12. *https://scispace.com/pdf/provenance-in-databases-past-current-and-future-ymbe17g99v.pdf*

Tian, F. (2016). *An agri-food supply chain traceability system for China based on RFID & blockchain technology.* In *13th International conference on service systems and service management (ICSSSM)* (pp. 1–6). IEEE. *https://doi.org/10.1109/ICSSSM.2016.7538424*

Weerakoon, B. S., & Chandrasiri, N. R. (2023). Knowledge and utilisation of information and communication technology among radiographers in a lower-middle-income country. *Radiography*, *29*(1), 227–233. *https://doi.org/10.1016/j.radi.2022.11.013*

Werder, K., Ramesh, B., & Zhang, R. (2022). Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems (TMIS)*, *13*(2), 1–23. *https://doi.org/10.1145/3503488*

World Health Organization (WHO). (2004). *Developing health management information systems: A practical guide for developing countries.* World Health Organization Regional Office for the Western Pacific. *https://iris.wpro.who.int/handle/10665.1/5498*.

World Health Organization (WHO). (2008). *Framework and standards for country health information systems* (2nd ed.). *https://www.who.int/healthinfo/country_monitoring_evaluation/who-hmn-framework-standards-chi.pdf*.

Zhang, J., Sun, J., & Stahl, J. N. (2003). PACS and web-based image distribution and display. *Computerized Medical Imaging and Graphics*, *27*(2–3), 197–206. *https://doi.org/10.1016/S0895-6111(02)00074-5*

Zhang, P., White, J., Schmidt, D. C., & Lenz, G. (2017). Blockchain technology use cases in healthcare. *Advances in Computers, 111*, 1–41. https:// *doi.org/10.1016/bs.adcom.2018.03.006*

Zhang, P., White, J., Schmidt, D. C., Lenz, G., & Rosenbloom, S. T. (2018). FHIRChain: Applying blockchain to securely and scalably share clinical data. *Computational and Structural Biotechnology Journal, 16*, 267–278. *https://doi.org/10.1016/j.csbj.2018.07.004*

CHAPTER 4

# STRUCTURING A DATA LAKE FOR THE MANAGEMENT OF SCIENTIFIC INFORMATION IN BRAZIL

**Washington Luís Ribeiro de Carvalho Segundo**
*Brazilian Institute of Information in Science and Technology (IBICT), Brazil.*
ORCID: *https://orcid.org/0000-0003-3635-9384*

**Fábio Lorensi do Canto**
*Central Library, Federal University of Santa Catarina, Brazil.*
ORCID: *https://orcid.org/0000-0002-8338-1931*

**Patrícia da Silva Neubert**
*Department Information Science, Federal University of Santa Catarina, Brazil.*
ORCID: *https://orcid.org/0000-0002-8909-1898*

**Adilson Luiz Pinto**
*Department Information Science, Pós-Design, Federal University of Santa Catarina, Brazil.*
ORCID: *https://orcid.org/0000-0002-4142-2061*
**Email:** *adilson.pinto@ufsc.br*

**Carlos Luis González-Valiente**
*Publications Department, Pro-Metrics, Tallinn, Estonia.*
ORCID: *https://orcid.org/0000-0002-1836-5257*

## ABSTRACT

The initial steps involved in the establishment of a data lake (Laguna) were delineated. This data lake was fed with structured data from the data ecosystem of the Brazilian Current Research Information System (BrCris). The data lake was developed to manage scientific information and aggregate this content into an accessible system. A substantial amount of data was collected and processed across five phases: (1) collection; (2) selection and separation; (3) transformation and connection; (4) organization, classification, and indexing; and (5) retrieval and visualization. The study utilized a range of data extraction methodologies on disparate platforms, employing SQL or API to facilitate the process. A set of scientific journals was identified through a process of stratification, with the highest percentage belonging to the A1 category. The initial integration of OpenAlex and DOAJ data was conducted, marking a significant milestone in the development of the platform. The author data were disambiguated and cross-checked by DOI to identify citing and cited authors. A comprehensive set of relevant data was obtained to facilitate the formulation of robust inferences, including the standardized number of journals by stratification, the integration between disparate databases such as OpenAlex and DOAJ, the ontological system employed to address the disassociation of authors, and the representation of the cited author before journals and future authorities.

**KEYWORDS**: data gap, data interoperability, scientific information management, academic databases

**HOW TO CITE**: Luís Ribeiro de Carvalho Segundo, W., Lorensi do Canto, F., Neubert, P. da S., Luiz Pinto, A., & González-Valiente, C. L. (2025). Structuring a data lake for the management of scientific information in Brazil. In A. Semeler (Ed.), Artificial Intelligence and Data Science Practices in Scientific Development, *Advanced Notes in Information Science, volume 8* (pp. 122-143). Pro-Metrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9331-4-5.112.

**COPYRIGHT**: © 2025 The author(s). This article is distributed under the terms of the CC BY-NC 4.0 license, which permits copying and redistribution of the material in any medium or format, adaptation, transformation, and building upon the material, provided that the license terms are followed.

## 1 INTRODUCTION

This research is supported by two projects: Laguna de Datos and the Brazilian Current Research Information System (BrCris; Pinto et al., 2022). These projects are overseen by the research group Brazilian Scientific Research Ecosystem Laboratory (LAEPECBR, in Portuguese). The objective of this initiative is to establish a data lake structure within Brazil, with the aim of supporting open data systems and ecosystems within the Brazilian Institute for Research in Science and Technology (IBICT, in Portuguese; Dias et al., 2022; Segundo et al., 2022). A significant challenge confronting Brazilian science and technology institutions pertains to the heterogeneity of the data recovered, characterized by a lack of cohesive structures. The construction of the IBICT data lake structure aims to address this challenge (Segundo & Sena, 2023). Concurrently, it aspires to function as a dependable repository for novel research data derived from the BrCris project. BrCris is an ecosystem that encompasses a comprehensive array of information and scientific findings. It employs sophisticated algorithms to derive indicators and metrics from recommendation systems, facilitating the identification of four distinct categories of specialists and specialties (Figure 1).

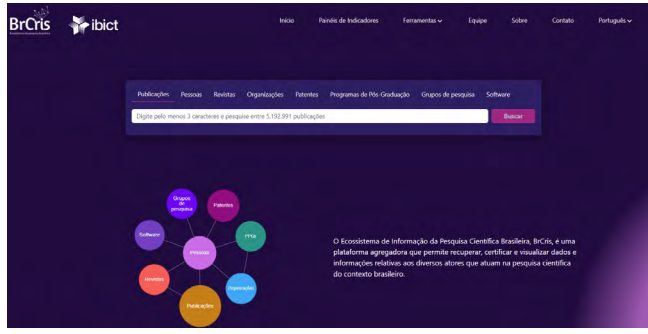


**Figure 1.** BrCris' webpage. **Source**. https://brcris.ibict.br

Laguna is defined as a system or repository of data stored in its natural format without processing (Ravat & Zhao, 2019). This system constitutes a unified repository of data derived from processed, statistical, and social systems. The objective of this system

is to transform, use, replicate, analyze, learn, and visualize the data in a manner that is accessible to all individuals involved in the systems, including system builders, operators, and users (Nargesian et al., 2019; Oliveira & Martins, 2022). The concept of context encompasses structured, semi-structured, and unstructured data, as well as image, audio, and video data. This type of data is referred to as "binary data" (Silberschatz et al., 2011). The data lake structure is predicated on six levels: (1) a management layer, (2) data access, (3) data collection tools, (4) various data repositories, (5) databases, and (6) a dashboard system so that the community has access to the contents (John & Misra, 2017; Valles-Coral et al., 2023). A data lake is defined as an extensive collection of datasets that can be stored in different systems (Giebler et al., 2019; Gontijo et al., 2021; Netto & Pinto, 2022). It is important to note that these data may be presented in a variety of formats and subject to change over time. The system generates autonomous operating systems, more appropriate reports, and predictive analyses for institutional needs. The data lake is a system that serves to structure a set of data according to its format specification, breakage of contents, content reformat, and data format. It also establishes dataset instances and connections, generates systems to qualify data and their contents, schedules, views, and accesses various data content (Coimbra & Dias, 2021; Segundo & Sena, 2023; Sousa & Shintaku, 2022). This particular data lake is employed by a diverse array of institutions, including companies, governments, and scientific-technological agencies.

The primary objective of its implementation is to ensure that the data are presented and stored in a scalable structure, with execution systems organized in clusters. These systems are required to process and store a substantial volume of data concurrently, in addition to executing these processes with open-source software and independently of file size. These systems are designed to collect potential data, identify user needs, monitor user behavior, detect fraud and data risks, manage marketing systems, analyze competition, and customize data demand. The objective of this initiative is to establish a data repository, designated as Laguna, which will serve as a foundation for the BrCris ecosystem. The specific objectives are as follows: (1) to incorporate statistical data, data aggregation APIs, and visualize certain scientific inferences, (2) to categorize scientific journals according to the Brazilian class identification model (Qualis/Capes), (3) to

identify open-access data crossover, (4) to perform author disambiguation for a more accurate representation of scientific data, and (5) to cross-reference the data by DOI to identify the degree of citations.

## 2 DESIGN AND METHODOLOGY

This research employed sophisticated computational methodologies for the management, structuring, and examination of information. This was done to obtain searchable, accessible, interoperable, and reusable datasets. The dataset has been meticulously collected, selected, transformed, and linked for the purpose of data processing. Subsequently, the data were methodically organized and indexed, and then retrieved and rendered visually on the BrCris platform. As a mining processing technique, six phases of data crossing were used: (1) understanding the scientific environment, (2) understanding of the data to be worked on, (3) preparation of these data, (4) mathematical modeling of these data, (5) evaluation of possible applicable models, and (6) generation of a data production system.

### *2.1 Data treatment model*

The data lake is defined as a collection of data that has been meticulously gathered from repositories and databases of recognized pertinence within the domains of science, technology, and innovation. The study focused on the direct engagement with the localization, accessibility, interoperability, and reuse of datasets, aligning with the FAIR Principles (Wilkinson et al., 2016). The following sources were consulted: OpenAlex, Wikidata, CrossRef, OpenCitations, OpenAIRE Research Graph, ISSN Portal, Latindex, DOAJ, Google Scholar Metrics, Plataformas Lattes and Sucupira, Oasisbr, and BDTD. The datasets obtained from these sources were then subjected to various analytical procedures, including advanced artificial intelligence techniques, real-time analytics, machine learning algorithms, dashboards, and visualizations. As the main results, we obtained:

1. *Collection*: We employed public APIs and search and extraction tools available in repositories and databases that fully or partially complied with the FAIR Principles. In addition, tools were developed for the purpose of extracting data from complex sources (do Carmo & da Silva Lemos, 2022). The utilized protocols encompassed the transmission and reception of messages through Representational State Transfer (REST) interfaces, with calls facilitated by HyperText Transfer Protocol (HTTP). This was done to obtain responses from documents in JSON format. Furthermore, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) protocol is employed, with HTTP calls and responses in the form of eXtensible Markup Language (XML) in accordance with various standards. In accordance with the tenets of the OAI—Dublin Core, we have obtained responses in more generic models, such as the Resource Description Framework (RDF), which exhibits a high degree of expressiveness.

2. *Selection and separation*: The subsequent filtration and categorization processes were then executed. Ancillary information for the collection was eliminated. The collected files were dismembered into the different types of entities described in their content, which, in this study, are referred to as "payloads."

3. *Transformation and connection*: Adaptations and validations were generated, as well as the establishment of relationships with records from other sources. A record obtained from source A exhibited a shared attribute with a record from source B, thereby enabling the establishment of a link between the two with a certain degree of reliability. The integration of the other record attributes resulted in the formation of a comprehensive, unified record. The process of duplication was eradicated.

4. *Organization, classification, and indexing*: The data that underwent classification served as the foundation for constructing search interfaces, web services, and visualization panels. The retrieval process involved the use of search facets in unstructured textual fields, encompassing full text, through actions such as tokenization and stemming.

5. *Retrieval and visualization*: The indicators were developed to facilitate the visualization of these metrics on the dashboards. The utilization of visualization tools resulted in the generation of collaboration networks, geospatial data, time series, and dynamic tabulation schemes. The semantic model employed in this study adhered to international standards. These systems were found to be compatible with those employed in other countries to achieve advanced levels of interoperability.

## 3 RESULTS

The results of the study are derived from the information sources outlined in the methodological framework. The SQL extractions used to identify journal priorities by the Brazilian graduate system were as follows:

### *3.1 Sucupira Platform to identify the journals and the potential of their indicators*

```
SELECT *
FROM (
SELECT sources.abbreviated_title, sources.alternate_titles,
        sources.apc_prices, sources.apc_usd, sources.cited_by_
        count, sources.country_code, sources.counts_by_year,
        sources.created_date, sources.display_name, sources.
        homepage_url, sources.host_organization, sources.host_
        organization_lineage, sources.host_organization_name,
        sources.id, sources.ids, sources.is_in_doaj, sources.is_oa,
        sources.issn_l, sources.publisher, sources.publisher_id,
        sources.societies, sources.summary_stats, sources.type,
        sources.updated_date, sources.works_api_url, sources.
        works_count, sources.x_concepts, qualis.issn, qualis.titu-
        lo, qualis.area_avaliacao, qualis.estrato
FROM laguna.sources, laguna.qualis
WHERE sources.type = 'journal'
AND array_contains(sources.issn, qualis.issn)
) LIMIT 5
```

### 3.2 *Google Metrics to identify scientific journals' h5 mean and median*

```
root
 |---title: string (nullable = true)
 |---issn: string (nullable = true)
 |---doi_example: string (nullable = true)
 |---title_gsm_2022: string (nullable = true)
 |---h5_2022: string (nullable = true)
 |---med_h5_2022: string (nullable = true)
 |---url_2022: string (nullable = true)
 |---title_gsm_2021: string (nullable = true)
 |---h5_2021: string (nullable = true)
 |---med_h5_2021: string (nullable = true)
 |---url_2021: string (nullable = true)
 |---title_gsm_2020: string (nullable = true)
 |---h5_2020: string (nullable = true)
 |---med_h5_2020: string (nullable = true)
 |---url_2020: string (nullable = true)
 |---title_gsm_2019: string (nullable = true)
 |---h5_2019: string (nullable = true)
 |---med_h5_2019: string (nullable = true)
 |---url_2019: string (nullable = true)
 |---title_gsm_2018: string (nullable = true)
 |---h5_2018: integer (nullable = true)
 |---med_h5_2018: integer (nullable = true)
 |---url_2018: string (nullable = true)
 |---title_gsm_2017: string (nullable = true)
 |---h5_2017: string (nullable = true)
 |---med_h5_2017: string (nullable = true)
 |---url_2017: string (nullable = true)
 |---title_gsm_2016: string (nullable = true)
 |---h5_2016: string (nullable = true)
 |---med_h5_2016: string (nullable = true)
 |---url_2016: string (nullable = true)
 |---title_gsm_2015: string (nullable = true)
 |---h5_2015: string (nullable = true)
 |---med_h5_2015: string (nullable = true)
 |---url_2015: string (nullable = true)
 |---title_gsm_2014: string (nullable = true)
 |---h5_2014: string (nullable = true)
```

```
|---med_h5_2014: string (nullable = true)
|---url_2014: string (nullable = true)
|---title_gsm_2013: string (nullable = true)
|---h5_2013: string (nullable = true)
|---med_h5_2013: string (nullable = true)
|---url_2013: string (nullable = true)
```

### 3.3 *OpenAlex to identify annual, 2-year, i10, and APC payments*

```
root
|---display_name: string (nullable = true)
|---issn: string (nullable = false)
|---issn_l: string (nullable = true)
|---2yr_mean_citedness: double (nullable = true)
|---h_index: long (nullable = true)
|---i10_index: long (nullable = true)
|---currency: string (nullable = false)
|---price: string (nullable = false)
|---apc_usd: long (nullable = true)
|---country_code: string (nullable = true)
```

With regard to the other databases, data recovery is achieved through extraction in CSV or JSON. As the primary source of information, we have OpenAlex, which is developing a journal recommendation system within BrCris, the final expression of the scientific index process. Given the substantial volume of data, we have implemented a data transformation process with Laguna, whereby we extract the raw data from their various sources and subsequently integrate them into BrCris. This approach was adopted to ascertain the most pertinent journals in the field. For instance, in the context of searching for a subject within the domain of journals (bibliometrics), there is often a multitude of variations in titles, which can pose a significant challenge. The Laguna mechanism facilitates this process, particularly through the incorporation of indicator options that are integrated into the system. Suppose that the search came up with 10 journal titles, such as *Em Questão, Informação & Sociedade: Estudos, Perspectivas em Ciência da Informação, Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação, Transinformação, Movimiento, Revista de*

*Pesquisa Cuidado é Fundamental Online*, *AtoZ: Novas Práticas em Informaçãoo e Conhecimento*, *Ciência e Saude Coletiva*, and *Revista de Saúde Pública*, and we need to know which one has the most publications on the subject. Subsequently, these metrics can be utilized in decision-making processes, with considerations given to the h5 mean or median, APC, and response time from the submission of the article to the receipt of an acceptance or rejection. The indices, which have been completed by Laguna and subsequently migrated to BrCris, can assist in determining a recommendation system. In the context of a system for journals, this technology can be utilized by editors to enhance the indexing of their periodicals. With regard to technological systems, such as patents, there is potential for enhanced accessibility for scientists.

The crux of the issue with the Laguna system is the absence of data transformation metrics, as the data are derived directly from the original sources. However, certain components of these sources are utilized, leading to the observed limitations. For instance, in the context of OpenAlex, the system has already developed a subset of the relevant metrics. The proposed solution involves the incorporation of recommendation systems. The study utilizes data from Google Scholar Metrics and the quality system of the Sucupira Platform. The ultimate objective of this study is to create indicators with the data recovered from information sources of magazines, theses, patents, and editorial content:

- *Scientific production*: We have some indices that can be managed, such as: (1) response time from submission to publication, (2) average H-5, (3) median H-5, (4) publication rate, (5) journal editorial committee—national or international, (6) inbreeding for Brazilian journals, and (7) languages of publication.

- *Theses*: We have the indices of (1) genealogy up to 11 levels, (2) regional orientation, (3) parents scientifically, (4) subject matter experts in guidance, and (5) thematic specialists in position shares.

- *Patents*: We can look at the system by (1) citations received, (2) quotes made, (3) concession and renewal, (4) patent family, (5) triadic, and (6) classification.

- **Editorial systems:** We have: (1) average H-5, (2) median H-5, (3) degree of endogamy, (4) production of doctors, (5) production of teachers, and (6) APC billing.

The visualization of this data is represented by priority indices within the BrCris dashboard, which is the output of all these indicators.

### *3.4 Scientific journal*

First, the records of scientific journals were cross-referenced using data extracted from OpenAlex and part of the contents of the Sucupira Platform, which is related to the Brazilian journal evaluation system Qualis/Capes. The enriched set of journal data was integrated into the BrCris platform, rendering it accessible for consultation and data visualization via a file. The file designates the journals as A1, A2, A3, A4, B1, B2, B3, B4, and C, thereby assigning them importance commensurate with that of the aforementioned journals. Figure 2 illustrates the distribution of journal classification strata within the Qualis system.

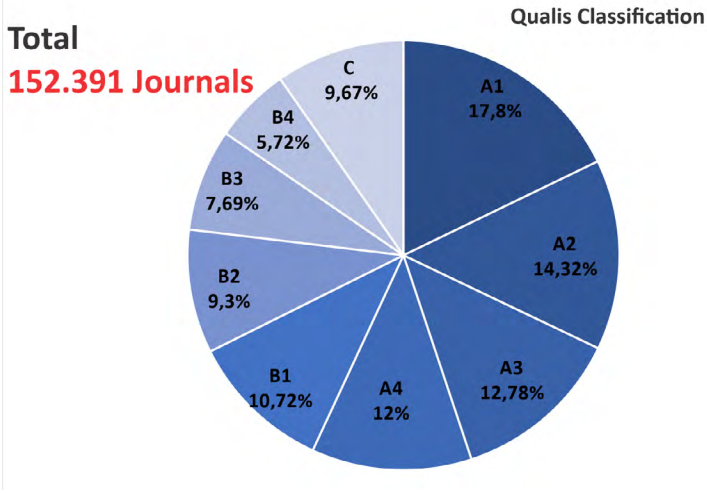


**Figure 2.** Dashboard of the journals in BrCris. **Source**. *https://brcris.ibict.br/vivo/revistas.*

### 3.5 *Open-access data crossing*

A cross-referencing process was conducted between data from OpenAlex and DOAJ and data from journals evaluated in the Qualis/Capes system. The variables that were identified during the course of this study included the percentage of open-access journals and the cost of APC by stratum and evaluation area. This dataset represents the initial phase of a cost foresight project that aims to develop a model of transformative agreements for Brazilian scientific production. A set of journals was selected for this study, and five impact indicators were previously calculated using data from open platforms. The indicators employed in this study include the 2yr_mean_citedness (average citations per article in 2 years), the h-index, and the i10 index, which were extracted from OpenAlex. Additionally, the h5 index and the h5 median were retrieved from Google Scholar Metrics. Subsequently, data analysis from open-access journals was carried out by cross-referencing data between the Sucupira, DOAJ, and OpenAlex platforms. The percentage of open access and the price of APC rates of the journals evaluated in Qualis/Capes (2017–2020) were identified, and the final data were quantified by stratum and evaluation area (Witt & Silva, 2022). Table 1 presents the mean APC price of journals by Qualis evaluation stratum. It has been observed that the APC price is elevated in journals of stratum A, particularly in stratum A1, which predominantly encompasses journals of heightened prestige and scientific rigor.

**Table 1.** APC price of Qualis journals by evaluation stratum.

| Qualis evaluation stratum | Average price of APC (USD) |
|---|---|
| A1 | 3,403.66 |
| A2 | 1,970.80 |
| A3 | 1,834.80 |
| A4 | 1,339.99 |
| B1 | 849.38 |
| B2 | 942.09 |
| B3 | 1,264.95 |
| B4 | 714.58 |
| C | 1,358.25 |

**Table 2.** Presents the mean APC price of journals by Qualis evaluation area.

| Evaluation areas | Average price of APC (USD) |
|---|---|
| Anthropology/Archeology | 2,675.68 |
| Astronomy/Physics | 2,669.72 |
| Biological Sciences II | 2,564.35 |
| Materials | 2,552.01 |
| Computation Science | 2,551.37 |
| Biological Sciences III | 2,547.07 |
| Biological Sciences I | 2,546.87 |
| Engineering IV | 2,518.90 |
| Chemical | 2,503.10 |
| Political Science and International Relations | 2,493.76 |
| Engineering III | 2,490.44 |
| Mathematics/Probability and Statistics | 2,490.40 |
| Pharmacy | 2,478.94 |
| Engineering II | 2,477.49 |
| Medicine I | 2,465.79 |
| Medicine II | 2,458.12 |
| Medicine III | 2,451.72 |
| Architecture, Urbanism, and Design | 2,444.74 |
| Economy | 2,422.27 |
| Engineering I | 2,392.18 |
| Collective Health | 2,370.94 |
| Biotechnology | 2,356.84 |
| Interdisciplinary | 2,355.33 |
| Psychology | 2,344.39 |
| Public and Business Administration | 2,305.44 |
| Physical Education | 2,300.72 |
| Geosciences | 2,287.76 |
| Nutrition | 2,260.39 |

| Evaluation areas | Average price of APC (USD) |
|---|---|
| Biodiversity | 2,249.48 |
| Food Science | 2,236.70 |
| Sociology | 2,221.95 |
| Art | 2,217.97 |
| Veterinary Medicine | 2,217.24 |
| Dentistry | 2,206.19 |
| Agricultural Sciences I | 2,197.00 |
| Geography | 2,194.22 |
| Law | 2,187.16 |
| Nursing | 2,184.24 |
| Communication and Information | 2,166.03 |
| Environmental Sciences | 2,162.37 |
| Linguistics and Literature | 2,118.91 |
| Teaching | 2,041.29 |
| History | 2,018.79 |
| Zootechnism/Fishing Resources | 2,000.49 |
| Urban and Region Planning/Demography | 1,871.05 |
| Philosophy | 1,808.08 |
| Education | 1,780.46 |
| Sciences of Religion and Theology | 1,719.75 |
| Social Service | 1,232.25 |

The observation revealed that the average price of APCs was notably higher in journals within the domains of Exact and Natural Sciences, Engineering, and Technology. Conversely, in the Human Sciences, Social Sciences, and Arts, the APC cost of journals is lower. Furthermore, a cross-referencing process was conducted between the h5 index of Google Scholar Metrics, OpenAIRE, and DOAJ. The initial dataset, pertaining to journals indexed in OpenAIRE, encompasses 15,366 titles, and the distribution of their h5-index values is illustrated in Figure 3.

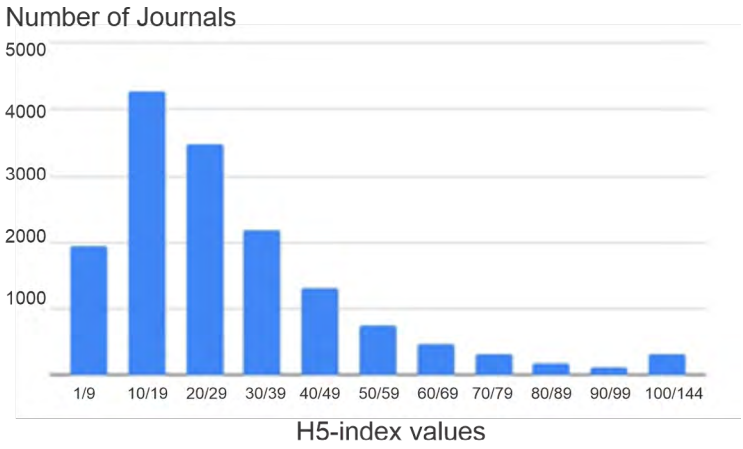


**Figure 3.** H5 index of journals indexed in OpenAIRE.

The second set included 9,722 open-access journals indexed in DOAJ (Vilas Boas et al., 2023), with ranges of h5-index values described in Figure 4. These data can be used to prepare studies on open access and transformative agreements in Brazil, as well as possible activities to be carried out in the project's subsequent phases.

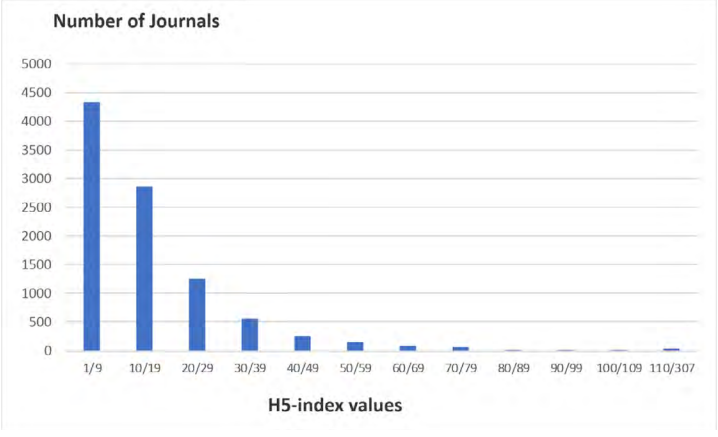


**Figure 4.** H5 index of journals indexed in DOAJ.

### 3.6 *Author disambiguation*

A significant challenge in the management of extensive collections of academic and scientific records pertains to the process of name disambiguation. To address this challenge, OpenAlex data are being integrated with the Lattes Platform, a Brazilian database that contains researchers' CVs (Mascarenhas et al., 2021). In the initial data cross-reference, approximately 100,000 authors with ORCID records in both sources were identified and extracted. This process facilitates the validation of author records and the expansion of the collaboration networks presented in BrCris, as illustrated in Figure 5.

**Figure 5.** Semantic visualization of authors in BrCris.

### *3.7 Data crossing by DOI*

Finally, OpenAlex data were extracted from nearly 5 million papers registered with DOIs in Lattes syllabi. The objective of this study was to analyze the existing data in order to identify a system of citations and cited references. The analysis entailed the identification of several key elements, including cross-citations of editors, the impact of journal citations, and the academic development of certain institutions within high-impact journals. The release of AWS credits from the call of the National Council for Scientific and Technological Development (CNPq, in Portuguese) initiated the training of project team members to configure Laguna's cloud infrastructure. The Laguna data were subsequently uploaded to the cloud server, and testing commenced using AWS processing and analysis tools. Among the tests carried out, the SageMaker notebook demonstrated a notable performance. This is a tool for developing machine learning models. The efficacy of the tool was ascertained through a citation analysis model that focused on scientific journals (Figure 6) and the subsequent generation of a network (Figure 7).

```
In [15]: entity_filter = 'concepts.id:'
         year_filter = 'publication_year:'
         year = '2022'

         # For demonstration purpose we are not going to do pagination
         ai_works = requests.get(openalex_api+'works?filter='+entity_filter+information_system_id+','+year_filter+year).json()
         df_ai_works = pd.DataFrame(ai_works['results'])
         df_ai_works
```

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9 | https://openalex.org/W4213072884 | https://doi.org/10.1016/j.techsoc.2022.101937 | Go cashless! Determinants of continuance inten... | Go cashless! Determinants of continuance inten... | 2022 | 2022-02-01 | {'openalex': 'https://openalex.org/W4213072884... | en |
| 10 | https://openalex.org/W4283378156 | https://doi.org/10.1016/j.techfore.2022.121562 | Analysis of the adoption of emergent technolog... | Analysis of the adoption of emergent technolog... | 2022 | 2022-05-01 | {'openalex': 'https://openalex.org/W4283378156... | en |
| 11 | https://openalex.org/W4210681325 | https://doi.org/10.1007/s12553-022-00639-w | Semantic interoperability in health records st... | Semantic interoperability in health records st... | 2022 | 2022-01-26 | {'openalex': 'https://openalex.org/W4210681325... | en |
| 12 | https://openalex.org/W4212841316 | https://doi.org/10.1016/j.autcon.2022.104171 | Design and implementation of a smart infrastru... | Design and implementation of a smart infrastru... | 2022 | 2022-04-01 | {'openalex': 'https://openalex.org/W4212841316... | en |
| 13 | https://openalex.org/W4220725585 | https://doi.org/10.3390/su14063508 | Antecedents and Impacts of Enterprise Resource... | Antecedents and Impacts of Enterprise Resource... | 2022 | 2022-03-16 | {'openalex': 'https://openalex.org/W4220725585... | en |
| 14 | https://openalex.org/W4205549242 | https://doi.org/10.1016/j.jclepro.2022.130557 | Green innovations, supply chain integration | Green innovations, supply chain integration | 2022 | 2022-03-01 | {'openalex': 'https://openalex.org | en |

**Figure 6.** Figure 6. Test of the scientific journal citation analysis model.

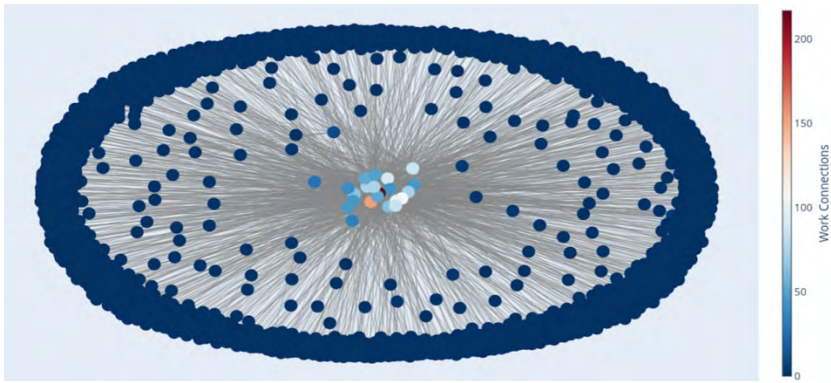


**Figure 7.** Network generated in testing the scientific journal citation analysis model.

## 4 CONCLUSIONS

In this study, we constructed the attributes constructor and proceeded to formulate a data lake model. The objectives that were previously outlined have been met, and the data serve as the foundation for the BrCris. The data lake is utilized for the processing of substantial datasets, encompassing both rationing and metrics. The development of certain APIs facilitated the extraction process. Consequently, the development of recommendation systems and the visualization of information in graph models and analytical graphs have become a reality. These initial findings may serve as a foundation for the development of novel research opportunities in the field of Brazilian ecosystems. The subsequent dataset is anticipated to be more extensive and dynamic, incorporating artificial intelligence and machine learning to facilitate the automated processing and indexing of data for aggregation within the BrCris framework. In contemplating the imminent future, it is anticipated that this technology will facilitate the generation and organization of data, thereby enabling the expeditious acquisition of information essential for effective decision-making. This initiative will be expanded to

encompass a broader range of disciplines, including patent data, technical production, and other scientific information types. The advantage of BrCris is that, regrettably, it is incapable of processing a significant amount of data, a capability that is possessed by Laguna. This phenomenon can be attributed to the fact that the systems in question—both hardware and software—were designed with this specific purpose in mind. The system is equipped with an AWS account, which facilitates the hosting of data across all levels. The Laguna serves to unify data in Brazilian science and technology, given that suppliers and services employ divergent methods. The primary function of the data lake is to develop the harmony of these data, a flexible model, and real-time machine learning. This approach ensures the accessibility and currency of the data, facilitating compatibility across a range of platforms, systems, programs, and tools.


## *Funding*

This study was supported by the Brazilian Institute of Information in Science and Technology.


## *Conflict of interest*

The author(s) declare that there is no conflict of interest.

## *Contribution statement*

**Conceptualization:** Washington Luís Ribeiro de Carvalho Segundo.
**Methodology and Data Curation:** Fábio Lorensi do Canto.
**Formal Analysis:** Patrícia da Silva Neubert.
**Writing – Original Draft:** Washington Luís Ribeiro de Carvalho Segundo, Adilson Luiz Pinto.
**Writing – Review & Editing:** Carlos Luis González-Valiente, Fábio Lorensi do Canto, Adilson Luiz Pinto.
**Supervision:** Washington Luís Ribeiro de Carvalho Segundo, Patrícia da Silva Neubert.

## REFERENCES


Coimbra, F. S., & Dias, T. M. R. (2021). Use of open data to analyze the publication of articles in scientific events. *Iberoamerican Journal of Science Measurement and Communication*, *1*(3), 1–13. *https://doi.org/10.47909/ijsmc.123*

Dias, T. M. R., Mena-Chalco, J. P., Segundo, W. L. R. C., Pinto, A. L., & Moreira, T. H. J. (2022). BrCris: Plataforma Para Integração, Análises E Visualização De Dados Técnicos-Científicos. *Informação & Informação*, *27*, 622–638. *https://doi.org/10.5433/1981-8920.2022v27n3p622*

do Carmo, D., & da Silva Lemos, D. L. (2022). Padrões de qualidade para dados e metadados endereçados a aplicações em ciência de dados. In *Advanced notes in information science* (vol. 2, pp. 161–170). ColNes Publishing. *https://doi.org/10.47909/anis.978-9916-9760-3-6.116*

Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2019). Leveraging the data lake: Current state and challenges. In C. Ordonez, I. Y. Song, G. Anderst-Kotsis, A. Tjoa, I. Khalil (Eds.), *Big Data analytics and knowledge discovery. DaWaK 2019. Lecture Notes in Computer Science* (p. 11708). Springer. *https://doi.org/10.1007/978-3-030-27520-4_13*

Gontijo, M. C. A., Hamanaka, R. Y., & de Araujo, R. F. (2021). Research data management: A bibliometric and altmetric study based on dimensions. *Iberoamerican Journal of Science Measurement and Communication*, *1*(3), 1–19. *https://doi.org/10.47909/ijsmc.120*

John, T., & Misra, P. (2017). *Data lake for enterprises: Lambda architecture for building enterprise data systems*. Packt Publishing.

Mascarenhas, H., Rodrigues Dias, T. M., & Dias, P. (2021). Academic mobility of doctoral students in Brazil: An analysis based on Lattes Platform. *Iberoamerican Journal of Science Measurement and Communication*, 1(3), 1–15. *https://doi.org/10.47909/ijsmc.53*

Nargesian, F., Zhu, E., Miller, R., & Pu, Q. (2019). Lake management: Challenges and opportunities. *Proceedings of the VLDB Endowment*, (2), 1986–1989. *https://doi.org/10.14778/3352063.3352116*

Netto, M. C. S., & Pinto, A. L. (2022). O silêncio dos dados diz muito, basta prestar atenção: breves experimentos sobre análise exploratória visual. In T. M. R. Dias (Ed.), *Informação, Dados e Tecnologia. Advanced Notes in Information Science* (vol. 2, pp. 15–23). ColNes Publishing. *https://doi.org/10.47909/anis.978-9916-9760-3-6.118*

Oliveira, L. F. R., & Martins, D. L. (2022). Coleta de dados para agregação de repositórios digitais: Entidades vinculadas à Secretaria Especial de Cultura do Brasil. In *Advanced Notes in Information Science* (vol. 2, pp. 171–181). ColNes Publishing. *https://doi.org/10.47909/anis.978-9916-9760-3-6.106*

Pinto, A. L., Segundo, W. L. R. C., Dias, T. M. R., Silva, V. S., Gomes, J., & Quoniam, L. M. (2022). Brazil Developing Current Research Information Systems (BrCRIS) as data sources for studies of research. *Iberoamerican Journal of Science Measurement and Communication*, *2*(1), 1–12. *https://doi.org/10.47909/ijsmc.135*

Ravat, F., & Zhao, Y. (2019). Data lakes: Trends and perspectives. In S. Hartmann, J. Küng, S. Chakravarthy, G. Anderst-Kotsis, A. Tjoa, & I. Khalil (Eds.), *Database and Expert Systems Applications. DEXA 2019. Lecture Notes in Computer Science* (pp. 11706, 304–313). Springer. *https://doi.org/10.1007/978-3-030-27615-7_23*

Segundo, W. L. R. C., & Sena, P. (2023, April 5–6). Laguna—FAIR research data infrastructure and open science support observatory [Conference session]. Expert Finder Systems, Coral Gables Miami.

Segundo, W., Dias, T. M., Moreira, T., Pinto, A. L., Silva, V., Gomes, J., Quoniam, L., Matas, L., Dias, A., & Schneider, J. (2022). Uma estratégia para coleta, integração e tratamento de dados científicos no contexto do BrCris. In T. M. R. Dias (Ed.), *Informação, Dados e Tecnologia. Advanced Notes in Information Science* (vol. 2, pp. 215–222). ColNes Publishing. *https://doi.org/10.47909/anis.978-9916-9760-3-6.117*

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2011). Relational database design. In *Database design concepts* (6th ed.). McGraw-Hill.

Sousa, R. P. M., & Shintaku, M. (2022). Política de privacidade de dados: observações relevantes para sua implementação. In T. M. R. Dias (Ed.), *Informação, Dados e Tecnologia. Advanced Notes in Information Science* (vol. 2, pp. 82–91). ColNes Publishing. *https://doi.org/10.47909/anis.978-9916-9760-3-6.112*

Valles-Coral, M., Injante, R., Hernández-Torres, E., Pinedo, L., Navarro-Cabrera, J. R., Salazar-Ramírez, L., Cárdenas-García, Á., & Huancaruna, E. (2023). Aggregation of institutional repositories for the analysis of the scientific performance of Peruvian universities. *Iberoamerican Journal of Science Measurement and Communication*, 3. *https://doi.org/10.47909/ijsmc.63*

Vilas Boas, R. F., Campos, F. F., Andrade, D. A. F., & Canto, F. L. (2023). Revistas científicas registradas no DOAJ: análise a partir do Índice H5. *BiblioCanto*, *9*(2), 100–115. *https://doi.org/10.21680/2447-7842.2023v9n21D33680*

Wilkinson, M. D., Dumontier, M., Jsbrand Jan Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. *https://doi.org/10.1038/sdata.2016.18*

Witt, A. S., & Silva, F. C. C. da. (2022). Analysis of citizen science in Brazil: A study of the projects registered in the Civis platform. *Iberoamerican Journal of Science Measurement and Communication*, *2*(3). *https://doi.org/10.47909/ijsmc.162*

CHAPTER 5

# DIALOGIC BRIDGES: VOICES BETWEEN IDEOLOGICAL FRONTIERS

**Manoel Camilo de Sousa Netto**
*Federal Police, International Cooperation Branch, Regional Superintendence in the State of Piauí, Brazil.*
ORCID: *https://orcid.org/0000-0002-7762-7958*
**Email:** *camilo.mcsn@pf.gov.br*

**Adilson Luiz Pinto**
*Department Information Science, Pós-Design, Federal University of Santa Catarina, Brazil.*
ORCID: *https://orcid.org/0000-0002-4142-2061*

## ABSTRACT

This study proposed a methodology for identifying deputies with the potential to establish a more effective articulation between disparate ideological blocs within a legislative body. This approach was predicated on network analysis. The network under consideration had, as its core, vote agreement among deputies, from which three metrics were calculated: bridge coefficient, betweenness centrality, and bridge centrality. Following a thorough descriptive and statistical analysis of the metrics in question, the 10 deputies with the most bridge centrality were identified. Bridge centrality was a metric that combined global structural influence with the connection between local communities. The study proffered a replicable methodology for identifying, solely on the basis of network structure, agents with a propensity to function as mediators in contexts of political fragmentation. This identification was a critical component of research focused on governability, the formation of coalitions, and the mediation of conflicts within multiparty systems.

**KEYWORDS**: social network analysis, bridge centrality, legislative agreement, political intermediation, inter-ideological articulation

**HOW TO CITE**: Camilo de Sousa Netto, M., & Pinto, A. L. (2025). Dialogic bridges: Voices between ideological frontiers. In A. Semeler (Ed.), Artificial Intelligence and Data Science Practices in Scientific Development, *Advanced Notes in Information Science, volume 8* (pp. 144-165). Pro-Metrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9331-4-5.113.

**COPYRIGHT**: © 2025 The author(s). This article is distributed under the terms of the CC BY-NC 4.0 license, which permits copying and redistribution of the material in any medium or format, adaptation, transformation, and building upon the material, provided that the license terms are followed.

## 1 INTRODUCTION

It is evident that the phenomenon of polarization has become a pervasive element in contemporary political discourse, exerting its influence across diverse national contexts. This trend is evident not only in formal parliamentary proceedings but also in the realm of social media interactions, underscoring its profound impact on the political landscape. The escalation of antagonistic discourses has led to a marked increase in the divide between groups with diametrically opposed viewpoints, a phenomenon that is often exacerbated by disinformation mechanisms and informational bubbles. The political landscape, once characterized by the use of dialogue and negotiation, transitioned to a model of confrontation. The empirical perception of an increase in polarization has been methodologically diagnosed in several scientific papers. A research study based on data extracted from the group of rounds from 2002 to 2018 of the Brazilian Electoral Study (ESEB) affirms that there is an increase in affective polarization in Brazil. This intensification of affective polarization became more pronounced in relation to leaders with clearer outlines as of 2018 (Fuks & Marques, 2022). In an environment characterized by political animosity, the identification of legislative actors who function as conduits between opposing ideological factions becomes a particularly salient issue. Deputies who possess the capacity to navigate between polarized groups often

serve as consensus builders, a role that is particularly essential in the context of legislative proposals that face challenges due to a lack of cross-party cooperation. The application of a moderate influence in the context of debates has the potential to facilitate the emergence of dialogue, even in the presence of significant political fragmentation. It has been posited that the recognition of these profiles may serve to enhance democratic deliberation and facilitate the overcoming of institutional impediments. Therefore, ascertaining the identity of these deputies emerges as a viable strategy to enhance the effectiveness and representation of the legislative process. Once identified, these individuals will be the primary targets in the effort to make the necessary talks viable.

The Brazilian legislative power offers a particularly fertile environment for this type of investigation, given its party plurality and the recurrent ideological fragmentation between deputies. According to Oliveira (2023), Brazil has been identified as the second most politically unstable nation, a typical case of high political fragmentation characterized by an absurd number of effective parties in its system. The legislative dynamics in Brazil are characterized by unstable coalitions, complex negotiations, and frequent reconfigurations in political alliances. This environment renders the actions of deputies capable of transiting between different ideological spectra even more challenging—as well as crucial. The Brazilian Legislative Branch can be regarded not only as an object of analysis but also as a strategic field of study to understand political mediation in polarized democracies. In this polarized fragmentation scenario, the deputies must act in concert to achieve a desired political objective, such as the approval of a law. To this end, they articulate, gather votes, persuade each other to vote, collaborate, and act as a cohesive group. This characterization enables the classification of this group of actors and their connections as an ideal raw material for social media analysis, the materialization of which may be implemented by the mathematical instrument known as a *graph*. Robins (2015) delineates a set of six compelling reasons for the incorporation of networks in the realm of social science research. Two of them, in particular, are the reasons that led to the selection of this theoretical basis for this study:

- **First reason:** The researcher wishes to study how individuals affect social structure.

- **Second reason:** If the individuals in certain social positions have different individual results.

The primary rationale pertains to the investigative scope of this study, as it is imperative to comprehend the legislative social structure and its interaction with the individual conduct of deputies to achieve the stipulated objective. The second reason is also pertinent, as it is of interest for the research to ascertain whether deputies occupying certain positions in the network are actors with the potential to collaborate in obtaining greater consensus. The objective of this study is to propose a methodology, grounded in social media analysis techniques, to identify deputies who serve as mediators between disparate ideological factions in the Brazilian Legislative Branch. The objective is to model parliamentary interactions as a network represented by a weighted, undirected graph. The purpose is to highlight actors whose structural positions turn them into individuals capable of establishing connections between distinct ideological communities. The utilization of enduring agreements predicated on votes has enabled the development of metrics that comprehensively gauge the capacity of intermediation and articulation among polarized groups. This analytical contribution aims to foster a more profound comprehension of public dynamics and identify potential cooperation agents in contexts marked by high fragmentation.

While the empirical focus of this study is the Brazilian Legislative Branch, the methodology adopted is of a general nature, allowing for application to any legislative system of democratic countries in which parliament maintains autonomy in the execution of its legislative functions. The proposed methods are not contingent upon specific institutional particularities, thereby enabling their adaptation to diverse national contexts. The analytical possibility is, therefore, valid in multiple representative democracies. The objective of this study is twofold: first, to understand a national reality, and second, to offer a methodological replicable tool to different nations.

## *1.1 Review of literature*

To identify the deputies who intermediate dialogues, it is necessary to understand them while also recognizing the network

actors that facilitate the flow of information between opposing groups. In the context of social media analysis, these actors are often identified through a specific metric: bridge centrality. This index quantifies the frequency with which an actor serves as an intermediary in the communication between other actors. It has been demonstrated that the extent to which an actor occupies a central position in the network of relationships among their peers is directly proportional to their influence in facilitating dialogue. First and foremost, it is necessary to establish the formal-theoretical basis of the bridge centrality metric. To that end, the following more basic concepts must be presented: degree of a node, betweenness centrality, and bridge coefficient. These concepts will be scrutinized as follows. One of the most essential concepts in graph theory, the degree of a node, is employed to calculate other metrics of a more complex nature. As elucidated by Caldarelli (2007), the degree of a node is defined as the quantity of edges connected to it. It is further clarified that the sum of all the degrees in a graph is equivalent to the number of its edges multiplied by two. This phenomenon occurs because each bond contributes twice in the degree counting process. Specifically, each bond contributes one unit for each of the vertices to which it connects. Given an adjacency matrix $A(n,n)$, the degree can be calculated through Equation (1).

*Equation 1*

$$d_i = \sum_{j=1,n} a_{i,j}$$

In turn, betweenness centrality is a measure of the global importance of a node that assesses the proportion of the shortest paths between all pairs of nodes that pass through the node of interest. According to Newman (2010), considering as the number of shorter paths between the vertices $s$ and $t$ that go through $v$ and considering $gst$ as the total number of shorter paths from $s$ to $t$, the betweenness centrality $c_B$ of the $v$ node in a general network is expressed by Equation (2).

*Equation 2*

$$C_B(v) = \sum_{st} \frac{n_{st}^v}{g_{st}}$$

The mathematical model under consideration posits that between two nodes ***s*** and ***t***, there may exist multiple paths of equal length, in addition to the single shortest path. In essence, the betweenness centrality of a node is the ratio of all shorter paths passing through that node. Hwang et al. (2006) have presented a semantic parallel that is intended to facilitate comprehension of the bridge coefficient. In the aforementioned article, the authors present a network as a simple electric circuit in which electrical current bottlenecks occur at the shorter degree edges. This phenomenon can be attributed to the reduced number of bonds in these nodes, resulting in a lower frequency of exits compared to higher-degree nodes. If the increase of a degree opens pathways for the flow, then its inverse would be some sort of "resistance," an obstacle to the course of information. Therefore, the bridge coefficient can be regarded as the underlying factor contributing to the observed resistance of a node to the sum of its neighbors' resistance. These regions, distinguished by their distinctiveness and dense connectivity, are linked by high-degree vertices, thereby facilitating the dissemination of information among groups. The bridge coefficient calculation $BC$ of a $v$ node is defined by Hwang et al. (2006) according to Equation (3).

*Equation 3*

$$B_c(v) = \frac{d(v)^{-1}}{\sum_{i \epsilon N(v)} \frac{1}{d(i)}}$$

where $d(v)-1$ is the inverse of the degree of node $v$, and $N(v)$ is the set of neighbors of node $v$.

It is evident that each of the components of the aforementioned trio of metrics possesses characteristics that delineate them as either local or global. While the "bridge degree" and "coefficient" are metrics local to the node, the "betweenness centrality" exhibits a more global relationship with the node. These characteristics support the hypothesis that the research's exploratory intention would be met by the deputies who serve as a bridge between different political communities within the Legislative Branch. Therefore, the research should not consider only the individual characteristics or the close political surrounding's characteristics, but also the global ones, from the political network to

which it belongs. To reunite both characteristics, the most adequate procedure was to adopt the metric named by the literature as "bridge centrality," whose value may characterize a vertex as a bridge node: a graph vertex situated between modules (or groups) that connect densely connected components. The value of this metric is obtained by multiplying the betweenness centrality by the bridge coefficient. This calculation considers both local characteristics stemming from the bridge coefficient and related to the node degree in relation to its neighbors, as well as global characteristics stemming from the betweenness centrality, which considers paths in a general manner. Specifically, the bridge centrality $C_p(v)$ for the node of interest $v$ is defined by Equation (4).

*Equation 4* $$C_p(v) = C_b(v) \times B_c(v)$$

## 2 METHODOLOGY

The research was conducted using a mixed approach, incorporating methods from both quantitative and qualitative perspectives in a non-exclusionary, complementary manner. Consequently, both numeric and descriptive data were utilized to provide a more comprehensive understanding of the phenomenon under study. The initial stage entailed data collection, mathematical formulation, and network creation. An exclusion criterion was devised for circumstantial concordances, and the bridge centrality calculation was executed. Finally, the deputies who exhibited the highest values for this metric were identified as the primary dialoguing actors of the House of Representatives for the year 2022, a year that was adopted by convention.

### *2.1 Data collection*

To conduct the studies, the research team obtained the open data of the House of Representatives from the aforementioned institution's website (Câmara dos Deputados, 2023). The data were available in groups named thematic collections. However, only

components of these records have been utilized in the research, namely deputies, ballots, and votes. A deputy is a representative elected by the people to exercise legislative mandate in a parliament, with the function of proposing, debating, and voting laws. The House of Representatives has a predetermined schedule that includes the matter of voting. As a formal procedure integral to the functions of the Legislative Branch, it is a process through which congressmen formally articulate their positions on various legislative proposals, including bills (law projects), amendments, and parliamentary motions. This event functions as a deliberative and decision-making mechanism within the scope of the branch. In the legislative context, voting signifies the individual expression of the parliament, which is executed during the act of casting a vote. This act is political and deliberative in nature, signifying agreement, disagreement, or abstention regarding a particular matter. Each vote contributes to the formation of the collective result of voting and reflects ideological, partisan, or strategic positioning. For the purposes of this research, a temporal window of one voting year and votes has been selected for the application of the data. The selected year was 2022. The data obtained are listed in Table 1.

**Table 1.** List of data used in the research and URL for obtaining said data. **Note**. Drafted by the authors.

| Data description | Source of data acquisition |
|---|---|
| Table of deputies | *https://dadosabertos.camara.leg.br/arquivos/deputados/xlsx/deputados.xlsx* |
| Votes (year: 2022) | *https://dadosabertos.camara.leg.br/arquivos/votacoes/xlsx/votacoes-2022.xlsx* |
| Voting table (year: 2022) | *https://dadosabertos.camara.leg.br/arquivos/votacoesVotos/xlsx/votacoesVotos-2022.xlsx* |

The concordance network under scrutiny in this study has been methodologically constructed using the three data sources presented above, as demonstrated in the following formalizations.

## 2.2 Mathematical formalization of the concordance network

From a methodological perspective, the establishment of a legislative social network, represented by a graph, necessitated the prior establishment of rules that govern the creation of vertices and edges. Each deputy has been represented by a vertex containing minimal data, such as a unique identification, name, and party affiliation. With respect to the edges' structures, their creation expressed the strength of concordance between two deputies (vertices). Upon participating in voting and sharing opinions through their votes, concerning any matter and intending to approve it (through "yes") or reject it (through "no"), those two congressmen may agree (through similar votes) or disagree (through different votes). The magnitude of the edge bonding between any two $A$ and $B$ deputies is directly proportional to the number of agreeing votes between them, as this results in greater proximity and interconnectedness between the deputies. This definition is part of the methodological conventions and seeks to contribute to the identification of the communities through metrics in later stages. Given that for each pair of identical votes of agreeing deputies, the edge that unites them receives the sum of one concordance unit (+1), the configuration necessitates the implementation of a weighted graph, with the weight attribute obligatorily linked to the edges. Furthermore, given that concordance is mutual and has no direction, the ideal graph is also undirected. The network structure can be formalized with support from Set Theory. Let $D$ be the set of deputies, defined as: $D = \{d_1, d_2, d_3, \ldots, d_n\}$. Let $V$ be the set of votings: $V = \{v_1, v_2, v_3, \ldots, v_n\}$. Let $R$ be the set of votes, where each element is a triple $(d_i, v_n, vote)$, representing that the deputy $d_i$ has voted in voting $v_n$. To consider the similarity between votes, we need to form unordered pairs of different deputies. Formally, the set of unordered pairs $\{d_i, d_j\}$, where $d_i \neq d_j$, is defined as Equation (5).

*Equation 5*

$$D' = \{ \{ d_i, d_j \} \mid d_i , d_j \in D \text{ and } d_i \neq d_j \}$$

It is imperative to note that the set $\{d_i, d_j\}$ is being considered, and consequently, $\{d_i, d_j\} = \{d_j, d_i\}$. For each pair $\{d_i, d_j\} \in D'$, a similarity function, denoted by $S$, is defined as the cardinality of the

set. That is to say, $S$ is the number of times that $d_i$ and $d_j$ both voted the same way in a given vote, as illustrated in Equation (6).

*Equation 6*

$$S(d_i, d_j) = |\{v \in V \mid (d_i, v, vote_i) \in R \text{ and } (dj, v, vote_j) \in R \;/\; vote_i = vote_j\}|$$

The resultant set of triples is expressed in the following format: $(d_i, d_j, w)$, where $w$ denotes the value of similarity $S(d_i, d_j)$. Given the unordered nature of the pairs, each pair is represented only once according to Equation (7).

*Equation 7*

$$T = \{ (d_i, d_j, S(d_i, d_j)) \;/\; \{d_i, d_j\} \in D' \}$$

### *2.3 Circumstantial concordances versus persistent concordances*

Following the establishment of the criteria for constructing the concordance network, it is imperative to prioritize resilient and enduring concordances over circumstantial ones in the analysis. This is due to the fact that only the former allow for the representation of ideological or behavioral bonds among agents with greater accuracy and stability over time. Momentary concordances, motivated by conjunctive factors or short-term specific interests, have the potential to distort the network's real structure by suggesting connections that do not reflect consistent alignments. A network created based on persistent concordances may reveal more robust patterns of cohesion and opposition, facilitating the identification of ideological communities and, consequently, the political actors that bond them with more analytical precision and longitudinal validity. In the parliamentary environment, the ideological alignment that endures is rooted in the consistent voting patterns exhibited over time. Therefore, given the configuration of the social network under consideration, the edges that exhibit higher relevance are those that have been assigned a greater weight than usual. These edges correspond to pairs of deputies who demonstrate consistent agreement, as

opposed to sporadic agreement. Conversely, when the weight is minimal, the exhibited bonds by the measured edges that bind the congressmen may have been attributable to transitional factors. The approach adopted in this study involves a simplification of the network, thereby facilitating the representation and interpretation of the most significant interactions during the visual and preliminary exploratory analyses. The rationale behind this phenomenon is that the social media presence of congressmen may be intricate, characterized by numerous nodes and edges, which can impede the identification of pictorial patterns. By emphasizing stronger connections, the visual representation of the resulting network becomes clearer and sparser, highlighting the cohesive groups formed within the parliament.

To maintain only the non-circumstantial concordances, the research has opted to preserve only the bonds with weight above the median. This will result in a network that emphasizes the strongly cohesive groups, offering a clearer and more objective view of the support relationships between pairs of deputies with similar ideological affinities. The maintenance of these stronger connections can be formulated mathematically as follows: suppose that $\boldsymbol{D}$ is the set of all the deputies represented in the network by vertices and that $\boldsymbol{E}$ is the set of edges, where each edge $\boldsymbol{e(d_i, d_j)}$ connects a pair of deputies $\boldsymbol{d_i}$ and $\boldsymbol{d_j}$. The weight $\boldsymbol{w(e)}$ of each set is defined as the sum of the concordance units between the deputies who are connected by the edge. Let be the median of the weight values $\boldsymbol{w(e)}$ for all edges in the set $\boldsymbol{E}$. Accordingly, we define the subgraph $\boldsymbol{G}$ as the set of all vertices $\boldsymbol{D}$ and edges $\boldsymbol{E'}$, where the weight of each edge $\boldsymbol{e'}$ is greater than according to Equation (8).

*Equation 8* $$G = \{ e \rightarrow E \mid w(e) > W_{median} \}$$

# 3 RESULTS

## *3.1 General characterization of the network built based on the methodology*

As delineated in the methodology, only enduring and non-circumstantial concordances have been retained. The median of the edge weights was subsequently calculated, and the edges with a lower or equal median weight were excluded. As illustrated in Table 2, the median value, the quantity of excluded edges, and the number of edges that remained subsequent to the exclusion of circumstantial concordances are indicated.

**Table 2.** Number of edges and vertices after the exclusion of circumstantial concordances. **Note**. Based on data from the House of Representatives, processed by the authors.

| Edges | Vertices | Median of weights (agreements) | Edges after deletion | Vertices after deletion |
|---|---|---|---|---|
| 152,464 | 554 | 153 | 75,854 | 501 |

## *3.2 Plotting the network as a graph*

The network plotting algorithm employed a directed force that was oriented in opposition to the plotting space communities that were ideologically opposed. Finally, the side and color of the vertices have been linked to the "bridge centrality." As demonstrated in Figure 1, larger and more pronounced vertices possess a higher metric value.

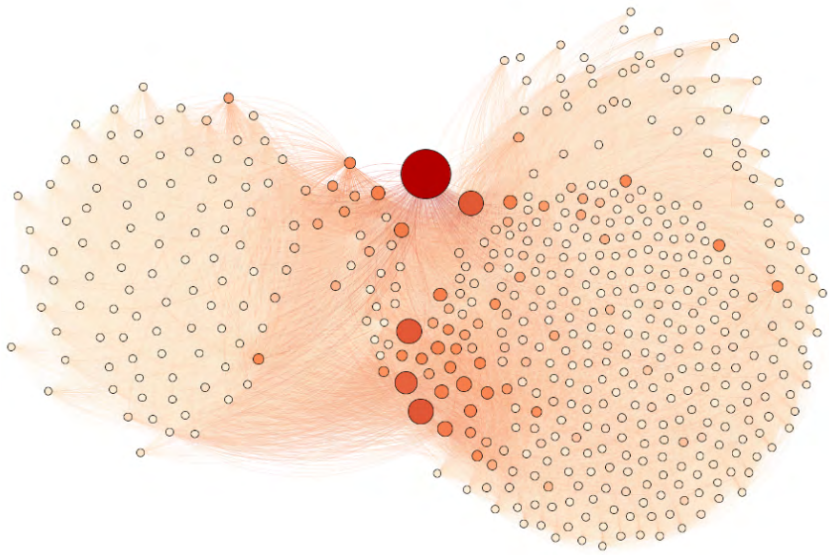

**Figure 1.** Plotting of the concordance network prepared following the methodology. **Note**. Drafted by the authors.

### *3.3 Descriptive statistics of the network's metrics*

As illustrated in Table 3, the measures of centrality and dispersion have been calculated for the nodes of the congressmen's social network. This offers a statistical characterization of the metrics network's distributions. The following central tendency values are included: the average and the median. The following dispersion values are included: the standard deviation, the first and third quartiles. The following forms are included: the asymmetry and the kurtosis for the betweenness centrality, the bridge coefficient, and the bridge centrality.

**Table 3.** Measures of centrality and dispersion of the congressmen's vertices. **Note**. Drafted by the authors

| | Betweenness centrality | Bridge coefficient | Bridge centrality |
|---|---|---|---|
| Average | 9.86275E-04 | 4.2986E-03 | 4.28418E-06 |
| Median | 2.4048E-04 | 4.1901E-03 | 1.0162E-06 |
| Standard deviation | 3.4901E-03 | 1.6735E-03 | 1.44275E-05 |
| Minimum | 0 | 2.3360E-05 | 0 |
| Maximum | 6.2152E-02 | 1.1556E-02 | 2.3662E-04 |
| First quartile | 8.81764E-05 | 3.2124E-03 | 3.2020E-07 |
| Third quartile | 6.7468E-04 | 5.4053E-03 | 2.8325E-06 |
| Asymmetry | 12.4510 | 2.4778E-01 | 10.5183 |
| Kurtosis | 197.0113 | 1.023463238 | 145.3964 |

### *3.4 Distribution of the centrality metrics*

The graphical analysis of centrality facilitates the identification of the distribution of values among congressmen and the reflection of network structural imbalances. For each metric, a histogram is presented, which shows the density of the frequency of values. A boxplot is also presented, which evidences medians, quartiles, and outliers. The histogram of the bridge coefficient reveals a relatively symmetric distribution, with slight positive asymmetry, as illustrated in Figure 2.

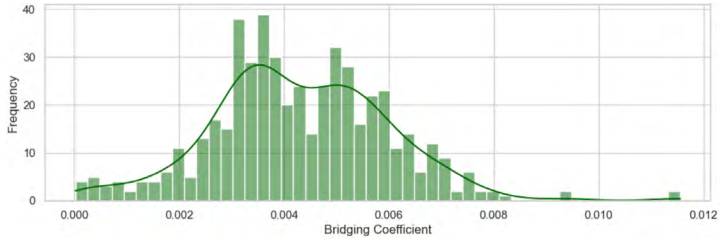


**Figure 2.** Distribution of frequency of the bridging coefficient. **Note**. Drafted by the authors.

The majority of the congressmen present exhibited values that approximate the mean, suggesting a less pronounced concentration of data in extremes. The boxplot graphic of the bridge coefficient confirms the presence of a moderate dispersion pattern, exhibiting minimal outliers, as illustrated in Figure 3.

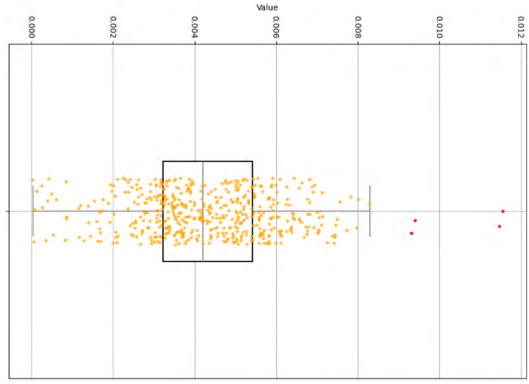


**Figure 3.** Boxplot of the bridging coefficient. **Note**. Drafted by the authors.

The linear scale employed in Figure 3 is adequate for effectively depicting the distribution, thereby reinforcing the homogeneity associated with this metric. Conversely, the histogram of betweenness centrality demonstrates pronounced asymmetry toward the right, with the majority of values concentrated around

zero and only a few instances exhibiting notably high values, as illustrated in Figure 4.

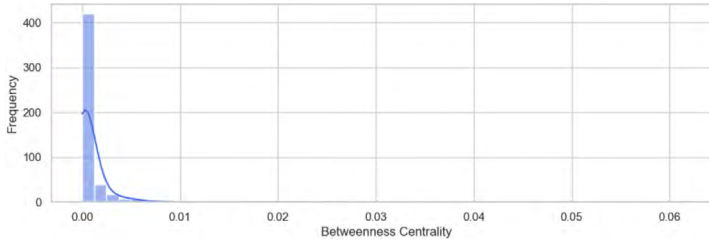


**Figure 4.** Distribution of betweenness centrality.
**Note**. Drafted by the authors.

This finding suggests that a significant proportion of congressmen do not occupy intermediary positions within the network. The betweenness centrality boxplot, with logarithmic scale, highlights a significant number of outliers above the third quartile, thereby reinforcing the concentration of the global centrality in a limited number of nodes. The utilization of a logarithmic scale for enhanced visualization is substantiated by the dispersed distribution, as illustrated in Figure 5.

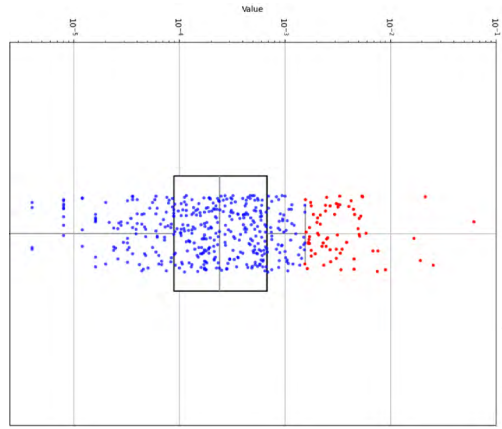


**Figure 5.** Distribution of the betweenness centrality.
**Note**. Drafted by the authors.

The distribution of bridging centrality exhibits a resemblance to the betweenness distribution, manifesting an accentuated asymmetry. The values are notably concentrated close to zero, with few congressmen assuming very high values, as illustrated by Figure 6.

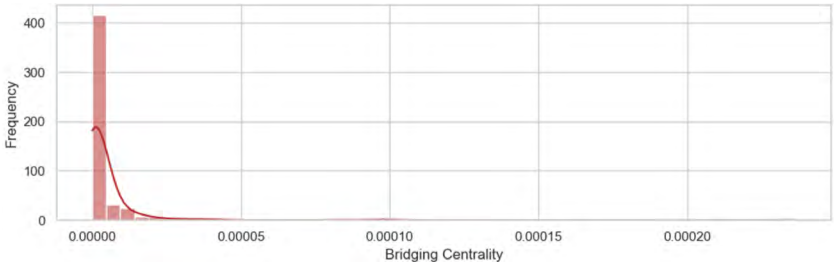


**Figure 6.** Distribution of bridging centrality.
**Note**. Drafted by the authors.

### *3.5 Congressmen with greater bridge centrality*

The results pertaining to the congressmen with greater bridge centrality are presented through their respective unique identifiers, as opposed to their names, as illustrated in Table 4.

**Table 4.** Congressmen with greater bridge centrality.
**Note**. Drafted by the authors.

| Unique identifier of the parliamentarian | Bridge centrality |
|---|---|
| 204433 | 2.3662E-04 |
| 74467 | 1.0051E-04 |
| 204466 | 9.7352E-05 |
| 204357 | 9.4965E-05 |

| Unique identifier of the parliamentarian | Bridge centrality |
|---|---|
| 160976 | 8.2423E-05 |
| 74471 | 4.3305E-05 |
| 74075 | 4.0648E-05 |
| 73460 | 3.9109E-05 |
| 90201 | 3.4778E-05 |
| 177282 | 3.3930E-05 |

The decision to omit the congressmen's names, replacing them with identifiers in Table 4, aims to redirect the analysis's emphasis toward the concordance network structural properties and the centrality metrics themselves. This approach serves to mitigate potential interpretation biases that might arise from immediate association with political figures. The analysis prioritized neutrality and objectivity, with the objective of facilitating the reader's comprehension of the network's patterns of interconnection and influence, unencumbered by preconceived notions about the individuals involved. Readers interested in identifying the congressmen corresponding to the identifiers may find the names associated with the aforementioned identifiers in an external repository. The comprehensive dataset pertaining to the parliamentary social network, encompassing the congressmen's identification numbers and their respective connections, is accessible to the public for download and reproduction. The vertices and edges files, which are sufficient for reproducing the social network, have been published in .csv format (tabular data separated by commas) and placed in the Zenodo repository. They can be accessed through the following DOI: *https://doi.org/10.5281/zenodo.15791268*.

## 4 DISCUSSION

The results obtained point to a significant inequality in the distribution of structural positions among the congressmen of the

legislative concordance network. The betweenness and bridge centrality metrics manifest strongly asymmetric distributions, with a substantial concentration of values near zero and the presence of outliers that are orders of magnitude greater. This pattern, characterized by elevated values of the high asymmetry index (12.45 and 10.52, respectively) and an extremely elevated kurtosis (197.01 and 145.39), supports the hypothesis that most congressmen occupy peripheral positions in the network, while a small number of individuals concentrate strategic connection roles. These values suggest not only a pronounced rightward skew in distributions but also a heightened density of values proximate to the mean, accompanied by more pronounced peaks. This phenomenon indicates the existence of a conspicuously influential structural elite. The betweenness centrality, by definition, considers the number of times a node acts as a bridge in geodesic paths among all the peers in the network. The global nature of the metric renders it particularly sensitive to the wide network structure, thereby justifying the strong asymmetry observed. A paucity of nodes has been observed that structurally bond communities or densely connected regions. The significant disparity between the mean (≈9.86 × 10-4) and the median (≈2.40 × 10-4) substantiates the presence of bias. Concurrently, the remarkably elevated kurtosis (197.01) indicates an exceptionally concentrated distribution, with pronounced peaks of atypical values, around the lower end of the spectrum. These findings suggest that the field of political network studies is characterized by the presence of influential minorities who possess high levels of articulation power.

In contrast, the bridge coefficient exhibited a considerably more symmetrical distribution (asymmetry of only 0.25) and a flatter distribution (kurtosis ≈1.02), suggesting a more homogeneous metric that is less influenced by extreme values. As it is a local metric, it evaluates if the direct connections of a node distribute themselves throughout different communities. Therefore, even congressmen who do not facilitate interactions on a global scale may exhibit moderate bridge coefficients if they establish bonds outside of their own blocs. This phenomenon elucidates the reduced dispersion, the approximation between the mean and the median, and the minimal presence of outliers. Consequently, the bridging centrality, which is derived from the weighted sum of its betweenness centrality and bridging coefficient, synthesizes both

local and global aspects of the network. The distribution exhibited a high degree of asymmetry (10.52) and a pronounced kurtosis (145.39), suggesting a significant influence of the global component on the metric. This finding indicates that, despite the limited number of congressmen who exhibit both high intermediation and intercommunity connections, this combination is functionally significant in identifying actors with the capacity to facilitate cross-cutting connections in contexts characterized by fragmentation and political polarization. In summary, the statistical and topological patterns observed indicate that the legislative concordance network is strongly hierarchized, with a limited number of congressmen occupying structural positions of intercommunity influence. The bridge centrality metric has been demonstrated to be an especially effective tool for identifying these actors, as it integrates both local and global structure dimensions. This ascertainment is pertinent to the comprehension of the parliament's internal dynamics and the development of replicable analytic methodologies that support studies about governability, consensus, and the formation of coalitions. The ensuing sections explore these implications in greater detail.

## 5 CONCLUSION

This study proposed a technical–analytical approach for the identification of congressmen with potential for intercommunity articulation in polarized legislative contexts. This approach entailed the modeling of social networks based on the concordances of votes. The utilization of conventional metrics (betweenness centrality) and derived metrics (bridge coefficient and centrality) permitted the isolation of structural attributes that differentiate peripheral positions of nodes occupying positions of strategic connection. From a methodological perspective, the combined utilization of global and local scope metrics was considered sufficient to characterize the congressmen with greater precision. These congressmen not only maintain bonds with different communities but also play a critical role in mediating the flow of decisions. The bridge centrality, upon synthesizing these dimensions, offers a composite metric that avoids the limits of unidimensional approaches and amplifies the sensitivity of structural analysis in political networks. Empirical evidence has demonstrated that

the asymmetry in the distribution of these metrics can be effectively diagnosed through the utilization of histograms and boxplots with a logarithmic scale. This approach enhances the interpretability of hierarchical structures and patterns of influence concentration. The implementation of these procedures in parliamentary networks enables the deduction of institutional positions based on empirically observed connectivity patterns and intermediation. This study offers a methodological contribution to the field of computational legislative analysis by proposing a replicable manner of localizing, from a network structure, the agents with greater potential of promoting ideological crossings. This identification is of particular pertinence for studies of governability, coalitions, and conflict mediation in multiparty decision environments.

## *Conflict of interest*

The authors declare that they have no conflict of interest.

## *Contribution statement*

**Manoel Camilo de Sousa Netto:** Writing – Review & Editing, Writing – Original Draft, Methodology, Conceptualization.
**Adilson Luiz Pinto:** Writing – Review & Editing, Supervision.

## *Statement of data consent*

This study generated vertex and edge data stored in the Zenodo repository: *https://doi.org/10.5281/zenodo.15791268*. The data used in this study are open and publicly available, and individual consent is not required for their use. The collection, processing, and analysis of these data were carried out in strict compliance with the General Data Protection Law (LGPD), Law No. 13,709/2018, and the Access to Information Law (LAI), Law No. 12,527/2011, ensuring legality and transparency in the use of information.

## REFERENCES


Caldarelli, G. (2007). *Scale free networks: Complex webs in nature and technology*. Oxford University Press.

Câmara dos Deputados. (2023). *Dados abertos*. *https://dadosabertos.camara.leg.br/*

Fuks, M., & Marques, P. H. (2022). Polarização e contexto: Medindo e explicando a polarização política no Brasil. *Opinião Pública, 28*(3), 560–593. *https://doi.org/10.1590/1807-01912022283560*

Hwang, W., Choy, Y.-R., Zhang, A., & Ramanathan, M. (2006). Bridging centrality: Identifying bridging nodes in scale-free networks. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (August 20–23, 2006), Philadelphia, PA, USA.

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.

Oliveira, A. N. C. de. (2023). Political institutions, electoral systems, and party stability in 40 democracies including Brazil. *Brazilian Political Science Review, 38*(2), Article e002. *https://doi.org/10.1590/1981-3821202300020004*

Robins, G. (2015). *Doing social network research: Network-based research design for social scientists*. Sage.

CHAPTER 6

# ANALYSIS OF PATENT PRODUCTION IN BRAZIL: A PERSPECTIVE FROM THE LATTES PLATFORM

**Dênis Leonardo Zaniro**
*Computer Department, Federal Institute of São Paulo (IFSP), Brazil.*
*Federal University of São Carlos (UFSCar), Brazil.*
ORCID: *https://orcid.org/0000-0003-2638-9264*

**Luc Quoniam**
*Federal University of São Carlos (UFSCar), Brazil.*
*Brazilian Institute of Information in Science and Technology (IBICT), Brazil.*
ORCID: *https://orcid.org/0000-0002-6333-6594*
**Email:** *quoniam.luc@gmail.com*

## ABSTRACT

**CONTEXT.** From the perspective of a nation's technological, economic, and social advancement, the assessment of technical yield was as crucial as the evaluation of scientific yield, with a particular emphasis on patents that safeguarded innovations and disseminated knowledge. Patents, therefore, represented a significant repository of technological information for various societal sectors, serving as a catalyst for innovation. In Brazil, a curriculum management platform known as Currículo Lattes enabled researchers to document their professional, scientific, and technical trajectories.
**OBJECTIVE.** Based on Currículo Lattes, this study provided an overview of the patent output of researchers working in Brazil by

analyzing the relationship between education level and the number of patents.
**METHOD.** The research was descriptive in nature and quantitative in its approach, employing statistical techniques to support its findings. The collection and analysis of data were facilitated by the development and execution of a set of algorithms within a computational framework.
**RESULTS.** The primary findings indicated that patent output increased over time, particularly in the past 25 years, and that researchers with doctoral degrees constituted the predominant proportion of inventors, thereby substantiating the study's hypotheses. Furthermore, by leveraging the mapping of patent output and the aggregation of open data on academic personnel from higher education and research institutions nationwide, it became feasible to conduct a comprehensive analysis of patent production and the collaborative dynamics that were inherent in each institutional context.
**CONCLUSIONS.** The study enabled the comprehension of the evolution of patent output by researchers in Brazil and the potential influence of education level on inventive activity in an unprecedented manner. The result achieved served as an important step toward the development of strategies and policies in the fields of science and technology, paving the way for new studies.

**KEYWORDS**: patents, Brazilian inventors, technical output, education level, Lattes platform

**HOW TO CITE**: Zaniro, D. L., & Quoniam, L. (2025). Analysis of patent production in Brazil: A perspective from the Lattes platform. In A. Semeler (Ed.), Artificial Intelligence and Data Science Practices in Scientific Development, *Advanced Notes in Information Science, volume 8* (pp. 166-190). Pro-Metrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9331-4-5.114.

**COPYRIGHT**: © 2025 The author(s). This article is distributed under the terms of the CC BY-NC 4.0 license, which permits copying and redistribution of the material in any medium or format, adaptation, transformation, and building upon the material, provided that the license terms are followed.

# 1 INTRODUCTION

The technological, economic, and social development of a nation and the world is directly related to the innovation capacity of organizations, governments, and various other productive and social arrangements (Schumpeter, 1980). Indeed, innovation is the prevailing term in contemporary discourse; it is situated at the intersection of research and industry, contingent on a dynamic process that integrates external and internal factors (Alvares & Itaborahy, 2021). A variety of external factors must be considered, including but not limited to: competitive pressure, technology transfer, the necessity of network cooperation, and compliance with regulations. The internal factors relevant to this discussion are typically associated with organizational culture, resource management, qualified human capital, investments in research and development, and other relevant factors. Furthermore, it is imperative to prioritize the quality achieved at all levels of the organization, in conjunction with sustainable development. Innovation encompasses a range of changes, including those in products, processes, or services. These changes are contingent upon the knowledge possessed by the economic and governmental sectors concerning the market, current technologies, social context, and their own internal structure (Bessant & Tidd, 2009). This knowledge pertains to technological information, also referred to as information for industry, and has been the focus of recent research in various academic disciplines (Braga & Simeão, 2018). The concept of "technological information" can be interpreted from multiple perspectives; therefore, there is a degree of variability in the literature regarding the scope of this concept. A definition that has been cited in numerous studies is provided by the International Federation for Information and Documentation (FID), a perspective that is further elaborated upon by Kariem (1990). According to FID, technological information is defined as "all knowledge of a technical, economic, market, managerial, social nature, etc., which, through its application, promotes progress in the form of improvement and innovation."

This definition is adequate for the objective of this study because, first, it emphasizes the multifaceted nature of technological information by establishing a connection with economic and social development, and second, it emphasizes its central role in promoting innovation. Consequently, patent documents

are considered a primary source of technological information on a global scale (Barroso et al., 2009; França, 1997; Mazieri et al., 2016; Quoniam et al., 2014). According to the World Intellectual Property Organization (WIPO) (2025a), a patent is defined as a legal instrument that is granted by the state with the purpose of protecting inventions within a given territory and over a given period. In addition, the patent system plays a significant role in innovation and, ultimately, in the economic and social development of a region or country from two main perspectives, according to Idris (2003), Pereira and Quoniam (2017), the WIPO Guide (WIPO, 2021), WIPO (2025a), and several other authors. On the one hand, the patent establishes a protection regime for the invention, thereby conferring upon the inventor the right to prevent others from commercially exploiting the patented object. On the other hand, a framework is established to facilitate the recovery of investments made in research and development by inventors and organizations. Conversely, the grant of a patent necessitates a comprehensive and detailed disclosure of its technical content, thereby establishing a foundation for the development of novel products, processes, or services. The optimal equilibrium between these two spheres of rights—the public and private—can serve as a catalyst for innovation, while concurrently ensuring economic viability.

Given the significance of technological information, particularly that derived from patent documents, as presented, it is imperative for any nation to devise methods to monitor and assess technical production, as indicated by patent filings. To accomplish this objective, it is imperative to determine the most suitable data source for extracting information concerning patent production within the country. In Brazil, a curriculum information system known as Currículo Lattes was developed and is currently administered by the National Council for Scientific and Technological Development (CNPq, 2023). This system is specifically designed for researchers operating within the Brazilian context. Currículo Lattes is a database that compiles information regarding the professional, scientific, and technical activities of researchers in Brazil. Consequently, this platform can function as a repository for patent production data, facilitating the identification of inventor researchers and enabling cross-references with other recorded data according to the researcher's informational requirements. The information declared in Currículo

Lattes is fundamental for the management of scientific and technological information and for the formulation of policies that foster the country's development in the fields of science, technology, and innovation (Silva & Smit, 2009). The Lattes curriculum database is also regarded as a reference for the approval of funding in research projects and activities (Oliveira et al., 2023).

However, as will be described below, studies in the literature analyzing data extracted from the Currículo Lattes database focus essentially on scientific production and collaboration. It is important to acknowledge that the analysis of patent data and inventor researchers through the Lattes platform and other databases is still an emerging research area, especially in Brazil. In light of the aforementioned context, the objective of this study is to provide a descriptive portrait of patent production in the country, the types of patents filed, the evolution in the number of filings over time, the education level of the inventors, and the collaborations involved in inventive activity. This portrait is based on data extracted from Currículo Lattes. Another contribution of the study, derived from the results achieved, is to enable the analysis of patent production and collaborations among responsible inventors within the context of each higher education and research institution through the implementation of a script. To this end, it is imperative to acquire a comprehensive list of the institution's personnel, adhering to specific guidelines, to facilitate the creation of an institutional inventor database for any higher education and research institution within the nation. The subsequent sections delineate the fundamental principles of patents and Currículo Lattes, which are indispensable for comprehending the investigative findings and scholarly discourse presented in this study.

## *1.1 Patents and inventions*

As previously stated, the primary functions of a patent can be categorized as follows: first, it serves to protect the patented object, and second, it serves to publish the technical information that constitutes the invention (WIPO, 2021). In Brazil, the National Institute of Industrial Property (INPI, in Portuguese, 2021) recognizes two forms of invention protection within the context of industrial property: invention patents and utility model patents.

Invention patents, as in many other countries, allow for the protection of new products or processes, that is, novel creations, and have a validity of 20 years counted from the patent filing date. Utility models are utilized to safeguard functional advancements or enhancements in the utilization or fabrication of particular practical objects, with a validity period spanning 15 years from the filing date. As is already known, both types of patents represent the legal instrument for protecting inventions. However, it is essential to distinguish the concept of patent from the concept of invention. A one-to-many relationship exists between the concepts of patent and invention. A patent is a legal document that grants its proprietor the exclusive right to practice the patented invention for a limited time. However, an invention may be covered by more than one patent, as the same invention can be patented in different regions and countries. In this particular instance, the invention is safeguarded by a patent family, as granted by the European Patent Office (EPO, 2017).

The concept of patent family will not be explored in depth here, but it is important to highlight that in Currículo Lattes, it is not possible to provide data on all members of a family when a given invention has been filed in different countries (CNPq, 2025). Consequently, even in instances where a family exists, typically only a single patent is documented in the inventor's curriculum. Conversely, patent documents that cite or are cited by a given patent declared on the Lattes platform are also not reportable in the system. In other words, there is a valuable patent ecosystem that cannot be identified solely through analysis of Currículo Lattes. This represents a notable limitation of the platform, and a methodology for identifying information on families, given and received citations, and other patent data involves the utilization of services furnished by international patent database platforms, such as the Espacenet database (Pereira & Quoniam, 2017). Espacenet, an online platform maintained by the EPO, is widely recognized as the world's most extensive patent database. It has been reported to aggregate patent data from over 100 countries (EPO, 2025). Currently, there are more than 150 million patent documents that are freely accessible through the interface or via the Web OPS (Open Patent Services) provided by the Espacenet system. This Application Programming Interface (API) is robust and is available in a free version, which enables the automation of data extraction and analysis processes (EPO, 2025). Information

of a technical, bibliographic, and legal nature can be extracted from patent documents maintained by Espacenet.

### *1.2 Currículo Lattes (Lattes curriculum platform)*

As previously stated, the Lattes curriculum platform, also known as Currículo Lattes, is an information system that is maintained by CNPq (2025) for the purpose of registering and consulting academic, scientific, and technological data of students, faculty, and researchers working in Brazil (Oliveira et al., 2023). The Currículo Lattes system was developed in 1999 (CNPq, 2023) and has been adopted as a standard consultation and analysis tool by the majority of funding agencies and institutions of education, research, and technology in the country (Mena-Chalco & César Júnior, 2009). To comprehend the life cycle of the Lattes platform, one must consider the three fundamental aspects outlined by Lane (2010): (1) the necessity to register and measure the nation's scientific activity was acknowledged, prompting the establishment of a collaborative community of federal agencies to design and develop the platform's infrastructure; (2) incentives were devised to motivate researchers and institutions to utilize the curriculum database effectively; and (3) a persistent identifier system, the Lattes ID, was implemented for researchers, thereby resolving conflicts caused by individuals with homonymous names. Currículo Lattes is an online system that is available free of charge to any individual who wishes to register their curriculum. To register, the user must first create an account (CNPq, 2025). In numerous academic and scientific contexts, researchers are obligated to register their data and maintain updated curricula (Bassoli, 2017). Consequently, in addition to its function as a curriculum database, Currículo Lattes serves as a substantial repository of scientific and technological information (Oliveira et al., 2023).

Given its importance to the scientific community, recent studies have investigated different categories of information declared in the curricula, such as activities, productions, projects, research lines, and fields of knowledge (Estácio et al., 2019). A comprehensive literature review was conducted to gather studies related to the platform's curriculum database over the past 20 years (from 2005 to 2025). The studies were subsequently

organized into three categories based on the object investigated in the Lattes platform: (1) the role of scientific networks and collaboration in the context of research, (2) the investigation of diseases and related phenomena, and (3) the analysis of data as a foundation for competitive and academic intelligence. The following studies were identified for Category 1: As indicated in the works of Balancieri et al. (2005), Dias et al. (2016), Dias and Moita (2018), Dias and Dias (2019), Dias et al. (2019), and Maruyama and Digiampietri (2021), the subject has been thoroughly researched. For Category 2, three studies have been identified: Magalhães et al. (2014), Sampaio et al. (2020), and Sobral et al. (2020) provide further insights into this phenomenon. Two studies are included in Category 3: Amaral et al. (2016) and Sarvo et al. (2023) provide further insights into this phenomenon. A substantial body of research has been dedicated to the evaluation of scientific production and collaboration in Brazil. This evaluation encompasses not only studies classified under Category 1, which prioritize the analysis of scientific data, but also those classified under Categories 2 and 3. These latter studies demonstrate a connection with the analysis of scientific production in Brazil, primarily through the utilization of Lattes curriculum as a data source.

Another study related to the Lattes platform, which is widely cited in the literature and serves as a basis for conducting different research, is the work by Mena-Chalco and Cesar Junior (2009). This study delineated the developmental process of the scriptLattes instrument, which facilitates the automated extraction and compilation of bibliographic, technical, artistic productions, advisories, and other pertinent information from researchers who have registered in the Currículo Lattes database. Specifically in the context of evaluating the technical production of researchers in Brazil through Currículo Lattes, only one study was found (Silva & Dias, 2023). This study offers the results of an analysis of patent production using the Lattes curriculum database, as well as the INPI (2021) and Espacenet (EPO, 2025) databases. However, the study considers only Brazilian patents (prefix BR), precluding the evaluation of the technical production declared by researchers registered in the Lattes platform in its entirety. Furthermore, the study by Silva and Dias (2023) does not demonstrate a correlation between patent production and the researchers' education level. These two aspects are fundamental to the proposal presented herein. The Lattes platform is a system

that enables researchers to freely provide their data, thereby ensuring the consistency and reliability of the data. Ensuring the consistency and reliability of the data is a challenge to be overcome. This discrepancy is further substantiated in the study by Brito et al. (2016), which examines the organizational shortcomings of information and the ambiguity surrounding guidelines for data filling or recovery. These problems, in general, have the potential to affect the accuracy, completeness, and credibility of data retrieval, thereby influencing the quality of the resulting information. Silva and Smit (2009) also emphasize that the platform has undergone significant advancements in recent years. To ensure the efficacy of this role, it is imperative to enhance the control and validation mechanisms of the declared information. This enhancement is necessary to prevent any compromise in information retrieval processes through the platform.

## 2 METHODOLOGY

This research employs a descriptive approach, as its objective is to characterize the scenario of patent production in Brazil by relating different variables in this context (Gil, 2010). To achieve this objective, a quantitative method was adopted, utilizing univariate and multivariate statistical techniques (Creswell, 2009). This study pertains to the domain of patentometrics (Hammarfelt, 2021), which aims to methodically analyze patent document data to discern various types of information, including technical, market-related, statistical, and others. As Nascimento and Speziali (2020) have demonstrated, this can contribute to research, development, and innovation. The methodology can be organized into five steps, as illustrated in Figure 1. Steps 1–4 were supported by the implementation of a set of algorithms, adhering to established software engineering practices (Pressman & Maxim, 2019). In Step 1, a comprehensive data extraction was conducted from the Lattes curriculum database through an API furnished by CNPq (2025), within the purview of an institutional agreement. The collection was conducted from late February to early March of 2025. The dataset encompasses approximately 9 million curricula from the Lattes platform, along with 115,258 patent records declared by researchers. In Step 2, an algorithm was implemented for the purposes of data cleansing, validation,

and deduplication, with a particular focus on the field designated for the storage of patent numbers, filing or publication numbers, as stipulated by the researcher in the curriculum. It is imperative to underscore the significance of this step, as it facilitates the reduction of inconsistencies and noise, thereby enhancing the quality of the data and, in turn, ensuring greater reliability in the analysis and interpretation of results.

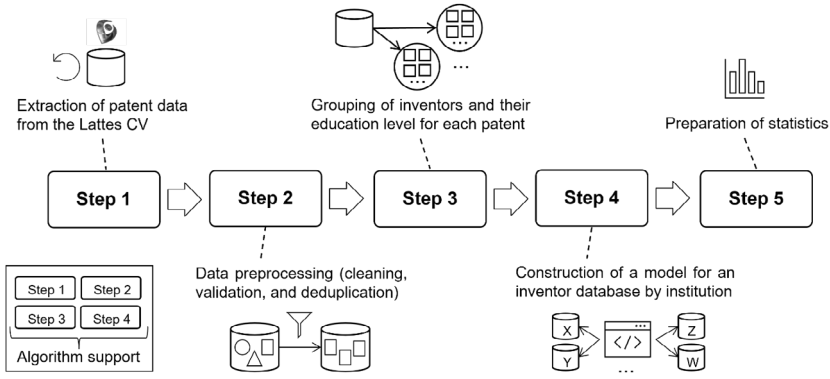


**Figure 1.** Research methodology steps. **Note**. Prepared by the authors.

The patent number underwent the following modifications for the purpose of cleaning: The stopwords were removed. These are special characters, spaces, and certain terms that frequently appear and are not part of the numbering; the invalid prefixes were removed. These are prefixes that do not correspond to a country prefix, and the kind codes were removed. The value of a kind code changes according to the patent status during its life cycle (published patent, granted patent, etc.). Consequently, the elimination of kind codes enhances the probability of identifying the same patent cited in the curricula of other researchers and locating this patent in additional databases, such as Espacenet. The subsequent data cleansing procedure entailed the classification of patent numbers as either valid or invalid. A valid patent number is one that contains at least four sequential digits (minimally, from the perspective of increasing the chance of finding that patent in other databases). The deduplication of patent records was conducted on the basis of cleaning and validation, with

the patent number serving as the criterion. Consequently, the number of unique patents declared throughout Currículo Lattes reached 65,173 (from a total of 115,258 patent declarations). The Lattes IDs of researchers were also deduplicated, revealing that 45,031 different curricula (researchers) declare at least one patent.

Subsequent to the deduplication of patent records, in Step 3, the inventor researchers were mapped and their educational attainment was ascertained for each unique patent. The categories of degrees and credentials include the following: PhD, master's degree, specialist, university graduate, and other (i.e., those who have not received a high school diploma or whose educational background is not specified). Consequently, it was feasible to obtain the count corresponding to each education level for each patent. In certain curricula, the same patent was declared multiple times. Consequently, the algorithm identified and disregarded these redundant declarations to ensure that the analysis of education levels was as faithful as possible to the Brazilian reality. Step 4 entailed the execution of a script that facilitated the generation of a database comprising inventors and their collaborative relationships for a specified education and research institution. To that end, it is imperative to obtain a list of institution staff members in CSV format, containing at least their full names. A cross-reference of the data from both datasets was conducted. The datasets in question consist of patent records (115,258) and the institution staff list. The name was designated as the pivot attribute in the cross-reference. For the purpose of comparison, the names are subjected to a series of processing steps. This processing involves the removal of stopwords, as well as the normalization of case sensitivity and diacritics. Finally, in Step 5, the data resulting from the preceding steps were synthesized according to each type of statistical analysis necessary to fulfill the study's objectives. In this study, only general statistics from the Lattes curriculum data are presented; however, given the results of Step 4, as described, it is possible to reproduce these statistics in specific contexts.

## 3 RESULTS AND DISCUSSION

The initial phenomenon investigated pertained to the temporal distribution of patent filings, extending from the inaugural filing

year documented in the Lattes curriculum (1900) to the present year (2025). This investigation was predicated on data from 65,173 patents declared within 45,031 curricula. The result of this investigation is presented in Figure 2. A notable aspect of the findings is that only one patent filing was identified for the years 1900, 1902, and 1923. There is also a significant 40-year gap with no patent records between 1923 and 1963. At this juncture, it is imperative to underscore that the Lattes curriculum is subject to perpetual refinement by researchers. Consequently, the analysis presented in this study is only a mere "snapshot" of the prevailing curricula during the data collection period. To supplement this analysis, a search was conducted in the Espacenet database (EPO, 2025) to attempt to find patents filed in Brazil (prefix BR) between 1900 and 1963. The system returned no records. The oldest Brazilian patent registered in Espacenet was published in 1965. It is imperative to acknowledge that the Lattes curriculum comprises records of patents filed in Brazil and numerous other countries. In an initial analysis, it is estimated that 80% of the patents declared in the curricula were filed in Brazil. From 2000 to 2020, the number of patent filings exhibited a consistent growth pattern. A decline was observed in 2021 and 2022 (e.g., from 2020 to 2021, the reduction was 16%), which may be attributable to the impact of the Coronavirus disease 2019 (COVID-19) pandemic. In 2023 and 2024, the number of filings exhibited an uptick, yet it remained below the quantity recorded in 2020, suggesting a partial recovery. The figure for 2025 is notably low, as it pertains exclusively to the months of January and February.

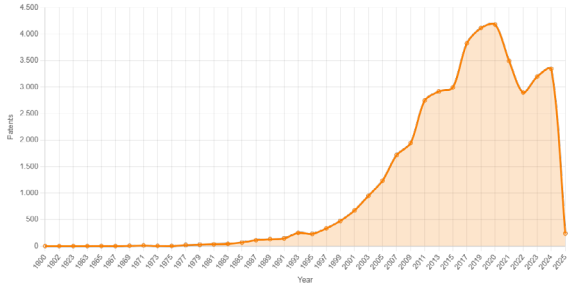


**Figure 2.** Number of patent applications over time.
**Note**. Prepared by the authors.

The ensuing analysis, as depicted in Figure 3, delves into the temporal progression of patent filings across the past three decades, distinguishing between various patent categories, including invention patents, utility models, and other types. This time frame was selected because the number of patents filed during this period corresponds to approximately 97% of the total patents filed and declared in the Lattes curriculum. Historically, the number of invention patents has consistently exceeded that of utility models, aligning with the trend observed in other countries that are among the 20 offices worldwide with the most patent filings (WIPO, 2025b). Examples of such nations include Japan, South Korea, India, and several European countries that have both forms of invention protection. From 1995 to 2020, the number of invention patents increased by more than 24-fold, while the number of utility models increased by 12-fold, approximately half of the aforementioned increase. The category designated as "Other" began to decrease in 2010 and was no longer reported in 2013, a development that may be attributable to a modification in the manner in which the patent type is documented on the Lattes platform.

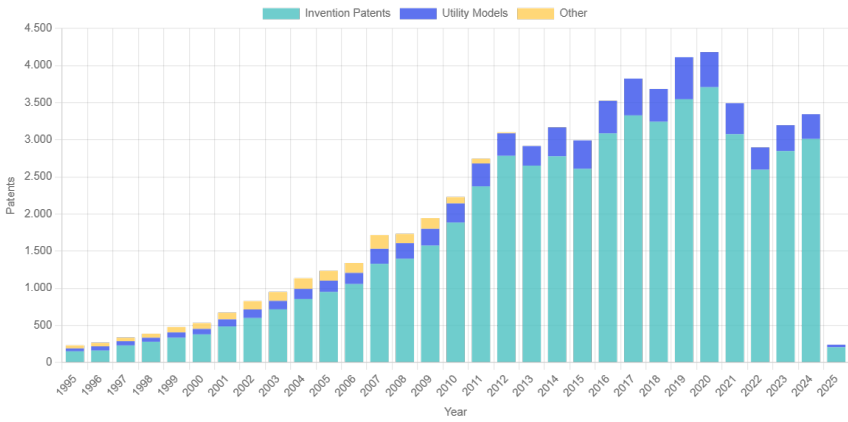


**Figure 3.** Patent applications by year and type over the past three decades. **Note**. Prepared by the authors.

Figure 4 presents the results of the investigation regarding patent production according to the inventors' education level, also over

the past three decades. A close examination of the data reveals that PhD holders constitute the predominant proportion of individuals filing patents on the Lattes platform. In addition to the data presented in Figure 4, it was determined that approximately 81% of the total patents declared in the Lattes curriculum have at least one PhD researcher listed as the inventor.

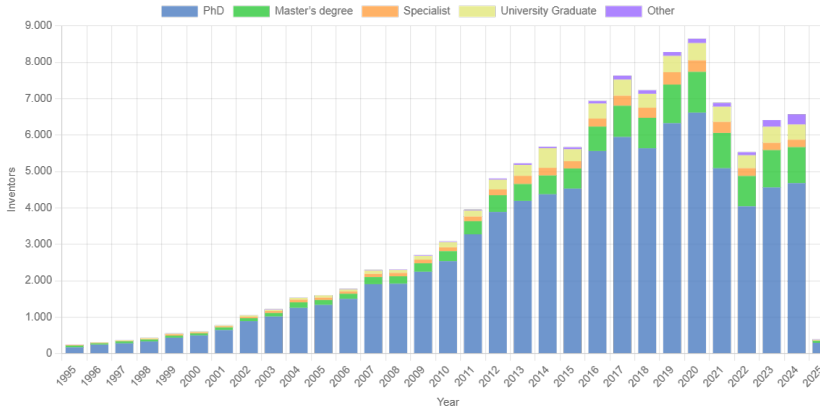


**Figure 4.** Patent applications by education level over the past three decades. **Note**. Prepared by the authors.

These figures indicate that the practice of patenting remains predominantly associated with postgraduate courses in Brazil, thereby substantiating the notion, as evidenced by several literature studies, that a significant proportion of patentable inventions emerge from scientific endeavors. To strengthen this hypothesis, please refer to Table 1, which shows the quantity of patents that involved collaborations between researchers with the same and different academic backgrounds. As demonstrated in Table 1, in 18,467 patents, there is collaboration of at least two PhD researchers; in 5,633 patents, collaboration between at least one PhD and one master's degree researcher occurs; and in 1,210 patents, collaboration between at least two master's researchers is present. Consequently, collaborations involving PhDs and master's degrees constitute 78% of all collaborations, and collectively,

researchers with doctoral and master's degrees are responsible for 90% of all patents.

**Table 1.** Collaborations among patent inventors by education level. **Note**. Prepared by the authors.

| | PhD | Master's degree | Specialist | Undergraduate | Other |
|---|---|---|---|---|---|
| **PhD** | 18,467 | 5,633 | 1,113 | 2,271 | 720 |
| **Master's degree** | 5,633 | 1,210 | 421 | 903 | 315 |
| **Specialist** | 1,113 | 421 | 149 | 293 | 93 |
| **Undergraduate** | 2,271 | 903 | 293 | 376 | 199 |
| **Other** | 720 | 315 | 93 | 199 | 98 |

While these figures are indicative of a potential causal relationship between education level and inventive capacity, it is imperative to carefully consider the implications of this result, given that an individual's education level is cumulative. Consequently, a researcher who currently holds a PhD may have filed patents prior to obtaining that degree. Consequently, further research is necessary to comprehensively investigate this association. Another potential analysis, based on the Lattes curriculum, involved a survey of educational institutions with the most patent filings, according to data reported by researchers. Table 2 presents the top 10 educational institutions with the highest number of filings, their respective states, and the number of patents filed by each institution. A total of 10 institutions were considered in the study. All of these institutions are public and predominantly federal. Furthermore, it was determined that 50% of the institutions are located in the Southeast region of Brazil. This region is notable for its large economy, which is the largest in the country. The aggregate number of patents filed by these 10 institutions constitutes nearly 20% of all patents declared in the Lattes curriculum. Consequently, the interplay between technological, scientific, and economic development is once again evident.

**Table 2.** The 10 universities with the most patents declared in the Lattes platform. **Note**. Prepared by the authors.

| University | State | Total of patents |
|---|---|---|
| Universidade Federal de Minas Gerais | Minas Gerais | 2,140 |
| Universidade Estadual de Campinas | São Paulo | 1,713 |
| Universidade de São Paulo | São Paulo | 1,512 |
| Universidade Federal da Paraíba | Paraíba | 1,185 |
| Universidade Federal de Campina Grande | Paraíba | 1,132 |
| Universidade Federal de Pernambuco | Pernambuco | 1,082 |
| Universidade Federal do Rio de Janeiro | Rio de Janeiro | 1,039 |
| Universidade Federal do Rio Grande do Sul | Rio Grande do Sul | 906 |
| Universidade Federal do Paraná | Paraná | 864 |
| Universidade Estadual Paulista Júlio de Mesquita Filho | São Paulo | 829 |

The data collection process revealed that, among a total of 27,083 patents, no applicant had been informed by the researcher. This finding offers a practical illustration of the challenges associated with data accuracy and completeness, as previously documented in various studies and further elaborated in this analysis. Consequently, alternative strategies must be implemented for precise data acquisition and analysis, typically involving queries to external databases. As previously delineated, an automated process was developed to generate a database of inventors for education and research institutions. This development was informed by data processing in the earlier stages of the study and the implementation of a script. The generation of the inventor database is contingent upon the procurement of a comprehensive roster of the institution's personnel, exemplified by the compendium furnished by the Brazilian Federal Government—Office of the Comptroller General (CGU, 2025). The process of data

matching between the patent records in the Lattes curriculum and the staff list is executed through a name comparison. Despite the inherent limitations of this approach, including the presence of homonymous names and divergent name specifications on the Lattes platform and the other source utilized, this component of the study can be regarded as an inaugural approach. Consequently, it is amenable to adjustments and enhancements, with the potential to facilitate a more comprehensive description of patent production and collaborative endeavors among inventors within diverse institutional contexts. Beyond enabling the measurement of its staff's technical production, the institutional inventor database has the capacity to facilitate the development of internal indicators, the identification of expertise in specific technological domains, and the mapping of partnerships. This contributes to the enhancement of technology transfer prospects. In the context of academic institutions, such as universities, the information extracted from the database can play a pivotal role in guiding various academic activities, including teaching, research, and outreach initiatives. This, in turn, can contribute to the cultivation of a robust intellectual property culture among faculty members, students, and other stakeholders within the institution.

## 4 CONCLUSION

The study provided a general overview of the patent production by researchers working in Brazil through the Currículo Lattes platform, revealing a relationship between education level and inventive activity, as well as how collaborations in the scientific context can be reflected in collaborations in the patent context. This patentometric study endeavors to facilitate a two-way street between scientific and technological information on one side and industry and government on the other. This movement is characterized by the dissemination of information stemming from technical production, particularly with regard to researchers' patents, which provides a foundation for decision-making processes at local, regional, and national levels. These decisions can subsequently result in funding and incentive policies aimed at transforming scientific and technological knowledge into innovation and development. Despite their close relationship, it is

crucial to draw distinctions between the terms "invention" and "innovation," as outlined by Schumpeter (1980) and subsequent scholars such as Mazieri et al. (2016). Schumpeter was the first to establish a link between these concepts and distinguish between different processes. According to Schumpeter, an invention can only be considered an innovation if it is introduced into the market context and produces some kind of economic or social effect.

Innovation is acknowledged to be contingent on numerous factors; however, its fundamental raw material is information and the knowledge derived from it. As thoroughly examined in this study and in other literature reviews, technological information, particularly that derived from patents, has emerged as a critical strategic asset for both organizations and nations. This strategic importance is particularly evident in the context of scientific, technological, and economic advancement. Despite this, there is a paucity of research in the literature proposing models or strategies for the effective conception, structuring, articulation, and utilization of technological information in its multiple dimensions and affecting different actors in the face of industrial, social, scientific, and sustainable demands. In this process, it is imperative to address the advancement of information and communication technologies, competitiveness, large volumes of data, green technologies, and other factors that influence and are influenced by technological information. This study unveils data and hypotheses that warrant investigation in subsequent studies, thereby facilitating not only diverse interpretations of the results but also the formulation of novel hypotheses concerning patent production, technological information, and its nexus with research, development, and innovation. As demonstrated in a variety of studies and as presented here, the Lattes platform occupies a central role in a research and innovation management system that facilitates connections between governments, educational and research institutions, funding agencies, and researchers. Consequently, it serves as a substantial repository of information for diverse research studies.

Conversely, as previously discussed, the Currículo Lattes is a self-declaratory system, thereby rendering the consistency and completeness of data contingent upon the manner in which researchers complete their curricula and the platform's inherent structure. In the context of patented inventions that pertain to families and involve citations, as illustrated in this study, a

substantial proportion of patent data cannot be disclosed exclusively through data extraction from the Lattes platform. In this context, this study constitutes a component of a broader project that involves the modeling of technological information as a complex whole. The objective of this endeavor is to serve as a foundation for establishing a connection between scientific research data, inventive processes, and industrial sectors. With regard to the patent dimension, there is ongoing work that represents the continuation of this study. This work is aimed at three major objectives. To accomplish these objectives, the work is grounded in an automated process of invention certification, consulting the Espacenet database services.

The initial objective is to generate a patent database from Currículo Lattes with accurate, precise, and consistent data, addressing the observed gaps in the platform and human errors in data entry. This result will also serve as a strategy to certify all patent declarations in each curriculum. This objective is consistent with specific objectives in Brazilian research related to open science, transparency, and higher credibility and quality of data. The second objective is to map all patent families to discover and systematize the countries and regions where each invention was filed, the complete set of International Patent Classification (IPC, 2025c) codes that allow identification of the technological fields of patents regardless of language, the languages in which the invention was written, as well as other information and analyses. At this stage, the study will transcend the patent concept toward a complete mapping of the inventions behind the patents declared in Currículo Lattes. The third objective is to map all patent citations given and received by the inventions declared in Currículo Lattes. From this objective, related technology information can be traced, thereby providing a broad technological basis for evaluating the impact and reach of inventions in various other studies. The culmination of this unparalleled endeavor is anticipated to represent a substantial advancement in the establishment of a technological information ecosystem that will foster research, development, and innovation in Brazil.

### *Acknowledgments*

We gratefully acknowledge the financial support provided by the Federal Institute of São Paulo (IFSP), Federal University of São Carlos (UFSCar), Brazilian Institute of Information in Science and Technology (IBICT), and Coordination for the Improvement of Higher Education Personnel (CAPES) for the development of this project. We also extend our sincere thanks to Professor Jesús P. Mena-Chalco for providing access to the dataset extracted from the Lattes platform.

### *Conflict of interest*

The authors declare that there is no conflict of interest related to this study.

### *Contribution statement*

**Dênis Leonardo Zaniro:** Data Curation, Investigation, Methodology, Software, Writing – Original Draft.
**Luc Quoniam:** Conceptualization, Methodology, Supervision, Validation, Writing – Review & Editing.

### *Statement of data consent*

The data generated during the development of this study are accessible at *https://tinyurl.com/zaniro*, under the terms of the CC BY-SA 4.0 license.

## REFERENCES

Alvares, L. M. A. de R., & Itaborahy, A. L. C. (Orgs.). (2021). *Os múltiplos cenários da informação tecnológica no Brasil do século XXI* (p. 474). Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). *https://labcotec.ibict.br/omp/index.php/edibict/catalog/view/280/290/1623*

Amaral, R. M., Brito, A. G. C., Rocha, K. G. S., Quoniam, L. M., & Faria, L. I. L. (2016). Panorama da inteligência competitiva no Brasil: os pesquisadores e a produção científica na plataforma Lattes. *Perspectivas em Ciência da Informação, 21*(4), 97–120. *https://doi.org/10.1590/1981-5344/2687*

Balancieri, R., Bovo, A. B., Kern, V. M., Pacheco, R. C. dos S., & Barcia, R. M. (2005). A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes. *Ciência Da Informação, 34*(1), 64–77. *https://doi.org/10.1590/S0100-19652005000100008*

Barroso, W., Quoniam, L., & Pacheco, E. (2009). Patents as technological information in Latin America. *World Patent Information, 31*(3), 207–215. *https://doi.org/10.1016/j.wpi.2008.11.006*

Bassoli, M. (2017). *Avaliação do Currículo Lattes como fonte de informação para construção de indicadores: O caso da UFSCar* [Dissertação de mestrado, Universidade Federal de São Carlos]. Repositório Institucional UFSCar.

Bessant, J., & Tidd, J. (2009). *Inovação e empreendedorismo*. Bookman.

Braga, T. E. N., & Simeão, E. L. M. S. (2018). A informação tecnológica no Brasil: Evolução da produção científica sobre o tema. *Informação & Sociedade, 28*(3). *https://doi.org/10.22478/ufpb.1809-4783.2018v28n3.41856*

Brito, A. G. C. de, Amaral, R. M. do, Faria, L. I. L. de, Quoniam, L. M., & Vieira, J. C. (2016). Visibilidade científica na Plataforma Lattes e Portal da Inovação. In: *Anais do XVII Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB)*. GT 07—Produção e Comunicação da Informação em Ciência, Tecnologia e Inovação.

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). (2023). *Plataforma Lattes*. Governo Federal. *https://www.gov.br/cnpq/pt-br/acesso-a-informacao/acoes-e-programas/plataforma-lattes*

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). (2025). *Plataforma Lattes*. *http://lattes.cnpq.br/*

Controladoria-Geral da União. (2025). *Portal da Transparência*. *https://portaldatransparencia.gov.br/*

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). SAGE Publications.

Dias, T. M. R., & Dias, P. M. (2019). Dados de pesquisa em acesso aberto: uma coleção de dados do conjunto de doutores cadastrados na Plataforma Lattes. *Ciência da Informação*, *48*(Suppl. 3), 518–519. *https://doi.org/10.18225/ci.inf.v48i3.4997*

Dias, T. M. R., & Moita, G. F. (2018). Um retrato da produção científica brasileira baseado em dados da plataforma Lattes. *Brazilian Journal of Information Science: Research Trends*, *12*(4), 62–74. *https://doi.org/10.36311/1981-1640.2018.v12n4.08.p62*

Dias, T. M. R., Moita, G. F., & Dias, P. M. (2016). Adoção da plataforma lattes como fonte de dados para caracterização de redes científicas. *Encontros Bibli: Revista eletrônica De Biblioteconomia E Ciência Da informação*, *21*(47), 16–26. *https://doi.org/10.5007/1518-2924.2016v21n47p16*

Dias, T. M. R., Moita, G. F., & Dias, P. M. (2019). Um estudo sobre a rede de colaboração científica dos pesquisadores brasileiros com currículos cadastrados na Plataforma Lattes. *Em Questão*, *25*(1), 164–188. *http://dx.doi.org/10.19132/1808-5245251.83-86*

Estácio, L. S. dos S., Viana, W. B., & Kern, V. M. (2019). O conhecimento sobre a Plataforma Lattes (CNPq) numa perspectiva sistêmica: Fundamentos e lacunas para estudos em Ciência da Informação. *Perspectivas em Gestão & Conhecimento*, *9*(1), 198–211. *https://doi.org/10.21714/2236-417X2019v9n1p198*

European Patent Office (EPO). (2017). *Patent families at the EPO*. *https://link.epo.org/web/Patent_Families_at_the_EPO_en.pdf*

European Patent Office (EPO). (2025). *Espacenet patent search*. *https://worldwide.espacenet.com/*

França, R. O. (1997). Patente como fonte de informação tecnológica. *Perspectivas em Ciência da Informação*, *2*(2), 131–140.

Gil, A. C. (2010). *Como elaborar projetos de pesquisa* (5th ed.). Atlas.

Hammarfelt, B. (2021). Linking science to technology: The "patent paper citation" and the rise of patentometrics in the 1980s. *Journal of Documentation*, *77*(6), 1413–1429. *https://doi.org/10.1108/JD-12-2020-0218*

Idris, K. (2003). *Intellectual property: A power tool for economic growth* (2nd ed.). World Intellectual Property Organization.

Instituto Nacional da Propriedade Industrial (INPI). (2021). *Manual básico para proteção por patentes de invenções, modelos de utilidade e certificados de adição* (versão jul-21). Ministério da Economia, Brasil.

Kariem, A. (1990). *FID Federation Internationale de information et de Documentation projects, programmes and problems: A select annotated bibliography* [Dissertação de mestrado, Aligarh Muslim University]. AMU.

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, *464*(7288), 488–489. *https://doi.org/10.1038/464488a*

Magalhães, J. L., Quoniam, L., Mena-Chalco, J. P., & Santos, A. (2014). Extração e tratamento de dados na base lattes para identificação de core competencies em dengue. *Informação & Informação*, *19*(3), 30–54. *https://doi.org/10.5433/1981-8920.2014v19n3p30*

Maruyama, W. T., & Digiampietri, L. A. (2021). Combinando agrupamento e classificação para a predição de coautorias na Plataforma Lattes. *Revista Brasileira de Computação Aplicada*, *13*(2), 48–57. *https://doi.org/10.5335/rbca.v13i2.12493*

Mazieri, M. R., Santos, A. M., & Quoniam, L. (2016). Inovação a partir das Informações de Patentes: Proposição de Modelo Open Source de Extração de Informações de Patentes (Crawler). *Revista Gestão & Tecnologia*, *16*(2), 52–75. *https://doi.org/10.20397/2177-6652/2016.v16i1.734*

Mena-Chalco, J. P., & César Júnior, R. M. (2009). scriptLattes: An opensource knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, *15*(4), 31–39. *https://doi.org/10.1007/BF03194511*

Nascimento, R. da S., & Speziali, M. G. (2020). PATENTOMETRIA: a utilização de dados contidos em patentes como mecanismo de análise da predominância tecnológica dos NITS. In: *IV Encontro Internacional de Gestão, Desenvolvimento e Inovação (EIGEDIN)*.

Oliveira, D. T. de, Rocha, L. de O., & Silva, P. N. (2023). Recuperação de informação no Currículo Lattes: proposição de requisitos aplicando técnicas de filtragem para recuperação da produção acadêmica. *Ciência Da Informação Em Revista*, *10*(1/3), 1–19. *https://doi.org/10.28998/cirev*.%y101-19

Pereira, S. de A., & Quoniam, L. (2017). Intellectual property and patent prospecting as a basis for knowledge and innovation—A study on mobile information technologies and virtual processes of communication and management. *RAI Revista de Administração e Inovação*, *14*(4), 301–310. *https://doi.org/10.1016/j.rai.2017.07.006*

Pressman, R. S., & Maxim, B. R. (2019). *Software engineering: A practitioner's approach* (9th ed.). McGraw-Hill Education.

Quoniam, L., Kniess, C. T., & Mazieri, M. R. (2014). A patente como objeto de pesquisa em Ciências da Informação e Comunicação. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação*, *19*(39), 243–268. *https://doi.org/10.5007/1518-2924.2014v19n39p243*

Sampaio, R. B., de Abreu Batista, A., Ferreira, B. S., Barreto, M. L. & Mena-Chalco, J. P. (2020). Scientometric analysis of research output from brazil in response to the Zika crisis using e-Lattes. *Journal of Data and Information Science*, *5*(4), 2020. 137–146. *https://doi.org/10.2478/jdis-2020-0038*

Sarvo, D. de O., Lozano, M. C., & Amaral, R. M. do. (2023). O uso de dados da plataforma lattes como fonte para inteligência acadêmica: análise de indicadores da produção científica das universidades públicas federais paulistas. *Informação & Informação*, *27*(3), 557–580. *https://doi.org/10.5433/1981-8920.2022v27n3p557*

Schumpeter, J. A. (1980). *The theory of economic development* (Originally published 1911). Transaction Publishers.

Silva, F. M., & Smit, J. W. (2009). Organização da informação em sistemas eletrônicos abertos de Informação Científica & Tecnológica: análise da Plataforma Lattes. *Perspectivas em Ciência da Informação*, *14*(1), 77–98. *https://doi.org/10.1590/S1413-99362009000100007*

Silva, R. R. da, & Dias, T. M. R. (2023). Analisando a produção técnica brasileira: uma abordagem considerando registros de patentes. *RICI: Revista Ibero-americana de Ciência da Informação*, *16*(1), 245–262. *https://doi.org/10.26512/rici.v16.n1.2023.47597*

Sobral, N. V., Duarte, Z., Santos, R. N. M. dos, & Mello, R. C. (2020). Redes de colaboração científica na produção de conhecimento em doenças tropicais negligenciadas no Brasil: estudo a partir da Plataforma Lattes do CNPq. *Encontros Bibli: Revista de Biblioteconomia e Ciência da Informação*, *25*(60), 1–27. *https://doi.org/10.5007/1518-2924.2020.e65476*

World Intellectual Property Organization (WIPO). (2021). *WIPO guide to using patent information*. *https://www.wipo.int/edocs/pubdocs/en/wipo-pub-rn2021-1e-en-wipo-guide-to-using-patent-information.pdf*

World Intellectual Property Organization (WIPO). (2025a). *Patents*. *https://www.wipo.int/en/web/patents*

World Intellectual Property Organization (WIPO). (2025b). *WIPO IP Statistics Data Center: Key indicators*. *https://www3.wipo.int/ipstats/key-search/indicator*

World Intellectual Property Organization (WIPO). (2025c). *IPC — International Patent Classification*. *https://www.wipo.int/en/web/classification-ipc*

CHAPTER 7

# A FRAMEWORK FOR COLLECTING, PROCESSING, AND ANALYZING SCIENTIFIC DATA ON SOCIAL MEDIA

**Thiago Magela Rodrigues Dias**
*Department Computer Science,* CEFET-MG, *Brazil.*
ORCID: *https://orcid.org/0000-0001-5057-9936*
**Email:** *thiagomagela@cefetmg.br*

**Rafael Gonçalo Ribeiro**
*Department Computer Science,* CEFET-MG, *Brazil.*
ORCID: *https://orcid.org/0000-0003-0646-6605*

**Patrícia Mascarenhas Dias**
*Department Computer Science,* UEMG, *Brazil.*
ORCID: *https://orcid.org/0000-0002-8448-6874*

## ABSTRACT

Given the increasing use of social media, it became imperative to understand the dissemination and discussion of scientific publications on these online platforms. The analysis of these data on interaction and circulation of scientific research was investigated in altmetrics studies and provided valuable information on how science was perceived and shared by the general public. The objective of this study was to propose a platform for the collection and analysis of social data related to scientific publications, with a focus on the video-sharing platform YouTube. By collecting data from YouTube, the platform sought to understand how scientific publications were disseminated and discussed on social

media. The Social4Science solution facilitated the acquisition of social data from YouTube and its correlation with scientific data from publications, enabling the analysis of multiple metrics. This methodological approach facilitated the identification of trends and patterns in discourse concerning scientific publications on social media. The findings indicated that the proposed platform held considerable promise in fostering a more profound comprehension of the interaction between science and the public. Furthermore, it had the potential to generate new avenues for future research in this domain. It was imperative to comprehend the manner in which scientific publications were received and discussed on social media platforms. This understanding was crucial for effective scientific communication and for fostering connections between the scientific community and the general public. The proposed platform contributed to this understanding, allowing researchers and professionals in the field to identify opportunities for engagement and develop effective strategies for scientific dissemination.

**KEYWORDS**: scientific production, social media, altmetrics, open data, bibliometrics

**HOW TO CITE**: Dias, T. M. R., Gonçalo Ribeiro, R., & Dias, P. M. (2025). A framework for collecting, processing, and analyzing scientific data on social media. In A. Semeler (Ed.), Artificial Intelligence and Data Science Practices in Scientific Development, *Advanced Notes in Information Science, volume 8* (pp. 191-207). ProMetrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9331-4-5.115.

**COPYRIGHT**: © 2025 The author(s). This article is distributed under the terms of the CC BY-NC 4.0 license, which permits copying and redistribution of the material in any medium or format, adaptation, transformation, and building upon the material, provided that the license terms are followed.

## 1 INTRODUCTION

The dissemination of scientific discoveries and the processes that facilitate it play a foundational role in the advancement of society and culture. The establishment of effective communication between the academic community and society is imperative, given that knowledge and research are intended to benefit society as a whole. Consequently, it is imperative that the method by which results are disseminated is congruent with the needs and expectations of the public. This is essential to establish a substantial and pertinent relationship between science and society in general (Neto, 2018). In this context, YouTube has demonstrated its significance as a primary platform for scientific dissemination on the Internet. As the most prominent global video-sharing platform, it encompasses a broad spectrum of content, encompassing numerous subjects and themes. In the context of Brazil, YouTube boasts a substantial and engaged audience, thereby establishing itself as a conducive platform for the dissemination of scientific knowledge (Da Fonseca & Bueno, 2021). The dissemination of scientific knowledge via YouTube has facilitated the creation of educational videos, debates, interviews, and practical demonstrations, thereby promoting interaction and dialogue between researchers and interested audiences. According to Reale and Martyniuk (2016), the dissemination of scientific knowledge via YouTube is an effective medium for democratizing scientific knowledge.

The analysis of scientific articles mentioned in YouTube videos offers the opportunity to collect a wide range of relevant data. This data may include the title of the scientific article, the names of the authors, the name of the journal in which the article was published, the year of publication, and the number of citations received, among other aspects of interest. This information is crucial for comprehending the interaction between the digital platform and scientific production, as well as for examining the impact and dissemination of scientific research on social media. The extraction of these data can facilitate the acquisition of insights regarding citation trends, the most frequently cited areas of research, and the subjects most prevalent in scientific videos on YouTube. This analysis also allows for the exploration of the connection between scientific dissemination and the academic framework, with the identification of the relevance and

influence of the scientific articles mentioned. A close examination of the citations in the videos reveals potential discrepancies between scientific research and its public dissemination, underscoring areas that merit heightened attention in the realm of scientific communication. Consequently, the extraction of data from scientific articles cited in YouTube videos signifies a pivotal approach to comprehending the nexus between scientific production and its dissemination in the digital domain. This contributes to a more comprehensive understanding of the propagation of scientific knowledge and its interactions with the general public. For instance, it is possible to identify emerging trends mentioned in the videos, thereby highlighting the most prominent and relevant topics in the realm of online scientific dissemination. Furthermore, it is possible to assess the influence of authors and journals, identifying those that are most mentioned and recognized on the platform.

This analysis facilitates a more profound comprehension of the dynamics underlying the dissemination of scientific research in the digital environment. A further critical component of this investigation entails the analysis of the relationship between the popularity of videos on YouTube and the number of citations received by the scientific articles mentioned in these videos. This correlation may reveal the influence of online videos on the dissemination and recognition of academic research. Comprehension of this relationship is essential for obtaining a comprehensive understanding of the interaction between scientific dissemination and the impact of research. A thorough analysis of the collected data may reveal gaps in scientific communication, indicating areas where there is a disconnect between scientific production and its online dissemination. These discrepancies may result in initiatives aimed at enhancing communication and public engagement, thereby fostering a more profound comprehension and esteem for scientific endeavors. In view of the aforementioned points, the objective of this study is to propose an innovative computational platform for the collection, processing, and analysis of scientific data on social media. It is imperative to underscore the utilization of the term "social media" in this context, as opposed to the term "digital social networks." The former term is more comprehensive, encompassing a broader array of online platforms that facilitate the creation and dissemination

of content, in addition to fostering interactions and connections between users.

The proposed platform, designated Social4Science, aims to address a significant gap in scientific research by providing an efficient tool to explore the vast universe of social media and understand how scientific information is disseminated, discussed, and perceived by the general public. The collection and analysis of data from this platform has enabled the acquisition of significant insights into science-related trends, patterns, and interactions. Moreover, the platform encompasses a wide range of functionalities that facilitate the identification of influencers, the analysis of the impact and relevance of scientific publications, the detection of emerging themes, and other significant analyses. In light of the aforementioned data, researchers will be in a position to make informed decisions, develop more effective dissemination strategies, and improve communication between the academic community and society. Consequently, Social4Science signifies a substantial methodology for investigating the possibilities of social media within the domain of scientific research. It provides a thorough and detailed perspective on the interactions between science and society, enhancing scientific communication, fostering inclusive dialogue, and establishing a robust connection between academia and the general public. The objective of the tool is to collect and analyze data from social media platforms, such as YouTube, with the aim of understanding how scientific publications are disseminated and discussed on these digital forums. Specifically, the objective of this study is to investigate the characteristics of videos published on YouTube that reference a Digital Object Identifier (DOI), with the aim of identifying relevant trends and patterns.

The objective of this study is to obtain results on how science is communicated and discussed in the online environment of YouTube. To this end, data will be collected from YouTube and analyzed using various techniques. By examining the characteristics of videos that contain DOIs, understanding of the manner in which scientific information is disseminated, the subjects that are addressed, and the manner in which the public engages with this content can be enhanced. Indicators of online attention have been a subject of discussion in the context of altmetric studies, which focus on understanding the social impact of research results on the social web (Araújo, 2020). These analyses can be

useful for researchers, journal editors, and other professionals involved in scientific dissemination, as they can help understand better how science is perceived and shared by the general public and to identify opportunities to increase the visibility of publications. Research that has been developed with these more contextual approaches is increasing in the extant literature, and it is indicative of the concern in the altmetric field to contribute to the deepening of the analysis and investigation of where and how articles are used by different communities that interact with them online (Araújo, 2020).

## 2 METHODOLOGY

This study employed the Altmetric portal, accessible via the Altmetric Explorer platform, as a tool to search for scientific publications that were cited in videos published on YouTube. The relationship between videos and scientific articles is established when a video mentions an article using the DOI, which is usually included in the video description. The utilization of the DOI as a unique identifier facilitates the precise linkage of a particular video to a corresponding scientific article. A search of the Altmetric portal for YouTube videos that mention DOIs revealed a dataset for analysis and study of the interactions between social media and scientific research. This approach of searching for references to scientific articles in YouTube videos using the DOI is an effective way to identify the presence and reach of science on this platform. The Altmetric Explorer platform offers resources that facilitate the collection and processing of data, enabling detailed analyses to be carried out on the characteristics of videos and citations of scientific articles. The entire data extraction process is initiated from a relationship extracted from Altmetric, containing a file with the video identifier and the DOI of a publication. The Social4Science platform receives this relationship as input and begins the entire data collection and analysis process, divided into two segments:

1. Social analysis: Data collection from YouTube videos.

2. Bibliometric analysis: Data collection from scientific articles.

The architectural design of the proposed platform is illustrated in Figure 1.

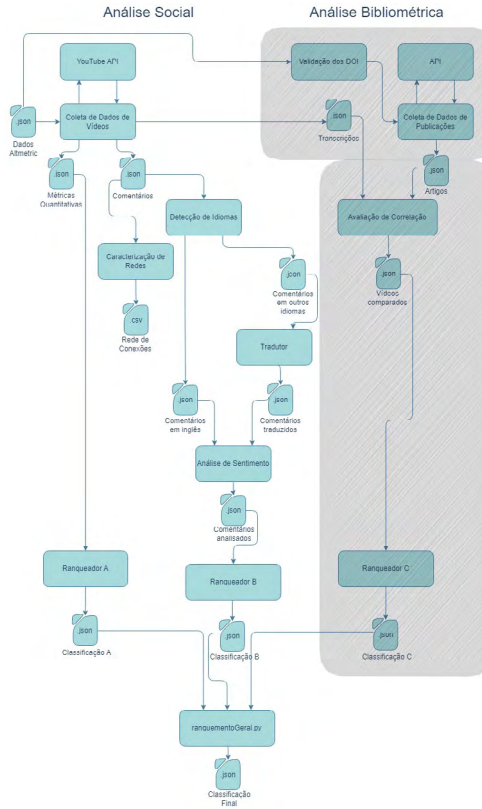


**Figure 1.** General architecture of the Social4Science platform.

The process of data collection and processing carried out by the platform commences with the input of a file provided by Altmetric, containing the video identifiers and the DOIs of the scientific articles. These DOIs are employed in the "bibliometric analysis" stage, while the video identifiers are employed in the "social analysis" stage. The subsequent "bibliometric analysis" stage entails the utilization of DOIs to procure pertinent information regarding the publications, including title, authors, journal

of publication, year of publication, and the number of citations received. These bibliometric data are essential for understanding the relevance and impact of the scientific articles mentioned in the YouTube videos. Consequently, the "social analysis" stage employs the video identifiers to investigate the social and interaction aspects associated with the videos that reference the scientific articles. This social analysis encompasses the identification of trends; the assessment of video popularity; the evaluation of user interactions, such as likes, shares, and comments; and the identification of relevant influencers or channels in scientific dissemination. The platform enables a comprehensive and in-depth approach to understanding the impact and dissemination of science on social media, especially on YouTube, by separating the bibliometric analysis and social analysis stages. The integration of bibliometric information and social data facilitates the acquisition of significant insights regarding the reception, discussion, and dissemination of scientific publications on this platform. This integration contributes to the advancement of scientific dissemination and the cultivation of a deeper understanding of the interactions between science and society.

In the context of "social analysis," video data are collected through the utilization of a publicly accessible YouTube Application Programming Interface (API), coinciding with the generation of specific data extracts. These metrics can be calculated and exported to other analysis and visualization tools, enabling further in-depth analysis. As a case in point, the sets comprising quantitative data from the videos, including the number of views, comments, and likes for each video, are emphasized. In addition, sets containing data from the channels in which the videos were published, the interaction networks identified from the comments on each video, extracts from the video descriptions, the transcriptions of each audio, and a set of standardized data in English from all the comments extracted are highlighted. In "bibliometric analysis," the set of DOIs is examined via API to ensure their validity. In the event that a DOI is found to be valid, the associated data are directed to the OpenAlex API, thereby facilitating the retrieval of information concerning the article in question. This includes details such as the article's title, authorship, year of publication, abstract, keywords, and the journal in which it was published, among other pertinent information. To

complement the data, a new request for the same DOI is sent to the OpenCitations API, retrieving the article's citations.

This comprehensive array of data is stored in data extracts that are also subject to analysis using various metrics implicit in the platform itself. These data extracts are made available in formats that can be imported by other analysis and visualization tools. Quantitative data play a fundamental role in the platform, allowing for different types of ranking and the analysis of correlations between social analytics and bibliometric analyses. These quantitative metrics offer valuable insights into the popularity, engagement, and reach of the videos and scientific publications referenced therein. Conversely, the datasets comprising textual information from videos, including titles, comments, descriptions, and transcripts, are correlated with the textual data from scientific publications, such as titles, abstracts, and keywords. In this context, correlation measures, such as the Levenshtein distance or the calculation of cosine similarity, are adopted to explore the relationships between the texts. The Levenshtein distance is a metric that calculates the difference between two sequences of characters, such as video titles and scientific publication titles. This measure enables the assessment of the thematic affinity or dissimilarity between the texts, thereby providing insights into the thematic proximity between the videos and the publications. The cosine similarity is a measure that quantifies the similarity between two-word vectors, such as the terms present in video comments and the keywords of scientific publications. This allows for the identification of semantic associations and relationships between the texts. The Social4Science platform employs correlation measures to reveal connections between the content of videos and scientific publications, identify thematic patterns, and explore how information is transmitted and discussed on social media.

Therefore, the integration of quantitative and textual data furnishes a thorough and enlightening analysis, enabling comprehension of the quantitative and textual dimensions implicated in the propagation and discourse of scientific publications on YouTube. To initiate the case study, a set containing 65,534 DOIs that had YouTube video citations at the time was collected from the Altmetric platform in March 2022. A series of verifications was conducted on a set of DOIs to ascertain the characteristics of scientific publications. A subsequent analysis of the publication

type revealed that the majority of the publications were classified as articles (94.9%), followed by a smaller proportion of books (3%) and book chapters (1%). It is also noteworthy that a total of 45 datasets were referenced.

## 3 RESULTS

Social analysis entails the examination of data derived from videos, including metrics such as the number of likes, views, and comments. Conversely, bibliometric analysis encompasses quantitative data derived from articles, including DOI validation, the number of citations received by other articles, and the number of videos that mention the article in question. The analysis of these data points enables the discernment of trends and patterns in discussions concerning scientific publications on social media platforms. For instance, it is possible to ascertain the most popular publications on these platforms, identify the topics that generate the most discussions, and determine the primary influencers in this context. Through bibliometric analysis, taking into account the date of data collection, the publication period of the articles mentioned in the videos was presented in chronological order (Figure 2).

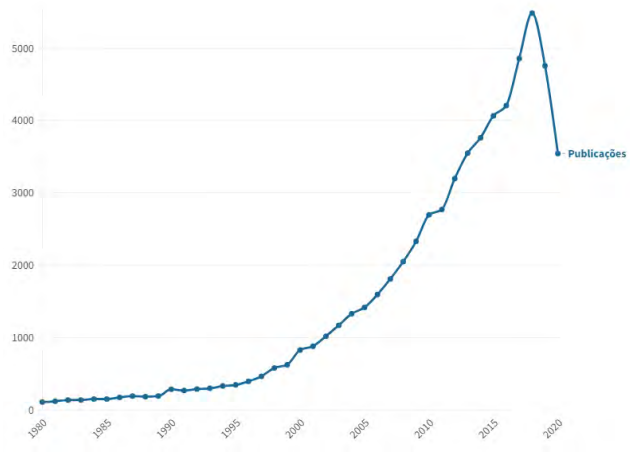


**Figure 2.** Publication period of the mentioned articles.

A subsequent analysis of the dataset revealed that the oldest article identified was published in 1980. A notable increase in the number of scientific articles is evident over time, with an even more pronounced trend from the year 2000 onwards. Concurrently, the utilization of DOI in scientific articles experienced a marked increase. A substantial surge in the number of scientific articles mentioned on YouTube was observed beginning in 2006, reaching its zenith in 2018. This growth can be attributed to a series of factors, such as advances in technology and research tools, broader access to scientific information, and increased collaboration between researchers on a global scale. The use of social media as a mechanism for disseminating research results has also played a significant role in this development. Additionally, the representativeness of the primary journals in which the articles were published could be ascertained. The objective of this analysis was to quantify the articles published in each journal, with a focus on identifying those that were most frequently mentioned in YouTube videos during the specified period (Figure 3).


**Figure 3.** Representation of journals in articles referenced in videos.

The following prestigious journals have been observed: *Nature*, *The American Journal of Clinical Nutrition*, *PLoS ONE*, *Nutrients*, *The Journal of Strength and Conditioning Research*, and *Science*, among others. These journals have gained international recognition for their editorial quality and the scientific rigor of their publications. It is noteworthy that specific domains of knowledge exhibit a higher prevalence of YouTube's utilization as a medium for the dissemination of scientific articles. This phenomenon can be attributed to several factors, including the nature of these areas, which are more readily transmitted through videos. It is important to note that certain regions may exhibit a heightened demand for direct and accessible communication, particularly in cases involving topics of public interest (Figure 4).

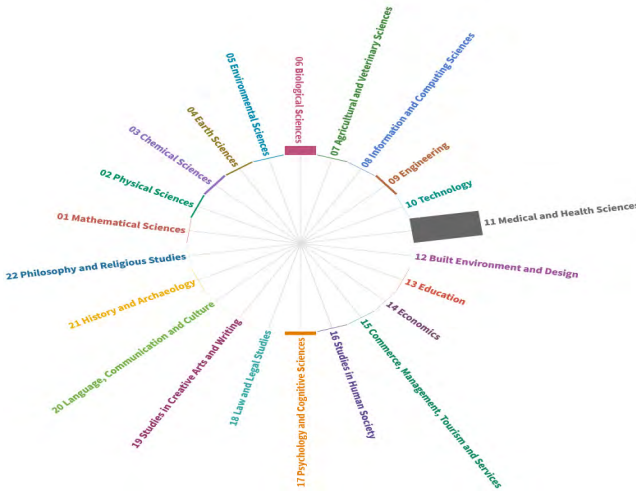


**Figure 4.** Predominant areas of the mentioned articles.

The utilization of YouTube as a scientific dissemination platform in these domains facilitates more dynamic and interactive communication, thereby providing a more engaging learning experience. Such videos may include a variety of content, such as practical demonstrations, interviews with experts, debates, and analyses of scientific articles, among other materials designed to

stimulate the interest and curiosity of the public. A close examination of the classification of the knowledge areas of the articles reveals a substantial concentration in two primary domains, as evidenced by the data collected. The analysis revealed that 69% of the articles were from the field of "Medical and Health Sciences," while 11.5% were from "Biological Sciences." These two areas encompass approximately 80% of the entire set of articles studied. This concentration in the domains of "medical and health sciences" is unsurprising, as these disciplines are inherently associated with human health and exert a substantial influence on individuals' lives. Scientific dissemination in these areas is of particular pertinence, as it facilitates the dissemination of crucial information regarding medical treatments, advancements in research, and disease prevention to the general public. The field of "biological sciences" also has a significant concentration of articles mentioned on YouTube. This phenomenon can be elucidated by the paramount significance of these studies in comprehending life and its biological entities. Topics related to the biological sciences, such as genetics, evolution, ecology, and biotechnology, have the potential to arouse the interest and curiosity of a wide audience. This, in turn, can contribute to the dissemination of content related to these subjects on YouTube. It is imperative to acknowledge that, despite the predominance of these two primary domains, other fields of knowledge are also represented in the articles that have been disseminated on YouTube, albeit to a more limited extent.

A number of additional bibliometric analyses were also conducted, including the validation of the DOIs cited, with the objective of verifying the authenticity of the publication. Subsequent to this stage, a collection and analysis of the data contained within the titles, abstracts, and keywords was undertaken, along with information regarding the number of citations of these articles by other publications. Additionally, data from the journals in which the publications were disseminated should be considered, including the impact factor and the Qualis. These data are systematically collected by public APIs from a variety of sources. The social analyses are predicated on information extracted from videos published on YouTube that include a DOI. This approach emphasizes the utilization of YouTube's public API for data retrieval, enabling the acquisition of information directly from the platform. The process of extracting data from YouTube is entirely

automated, commencing with the initial list provided, in which the video identifiers are extracted and the requests are made to the YouTube API. A subsequent analysis of the collected data revealed that the videos were classified into categories. These categories refer to the channels in which these videos are published (Figure 5).

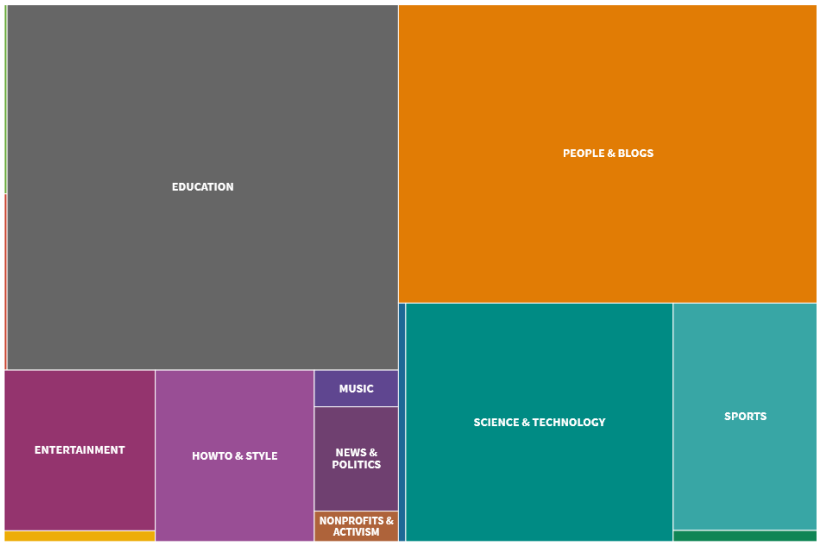


**Figure 5.** Category of channels where videos are published.

A survey of video content reveals that the majority of channels that present videos with DOIs primarily fall into the categories of "education," "people and blogs," and "science and technology." One hypothesis for the greater representation of the "education" category may be related to classes or dissemination of study results. To achieve a more profound comprehension of this categorization and its repercussions, it is imperative to undertake a thorough examination of the representation of these channels, taking into account the number of subscribers, the quantity of videos published, and the date the channel was registered on YouTube. This information can yield additional results and further enrich the analyses. In addition, the data from "social analysis" indicate

that comment networks are established through the developed platform. A subsequent analysis of the comments associated with each video is then conducted, with the objective of identifying the connections between the various channels. Therefore, a set of comments can be utilized to facilitate an analysis of the interactions between channels, with consideration given to the comments that are made or received by them. This network analysis approach facilitates the acquisition of significant findings regarding the dynamics of interactions between channels in the context of the comments.

In addition to the characterized networks, several other quantitative analyses are performed that aggregate relevant information. A comprehensive set of data, including metrics such as the number of views, comments, likes, duration, and language of the videos, is considered. These metrics offer valuable insights into the reach, engagement, and characteristics of the videos, thereby facilitating a more comprehensive understanding of their relevance and impact on the platform. The content of the videos is also the object of study, covering elements such as the title, description, and audio transcription. These elements are of paramount importance, as they facilitate numerous analyses aimed at correlating the content of the videos with the data from the scientific articles, which are also collected, such as title, abstract, and keywords. These analyses facilitate a more profound comprehension of the subjects addressed in the videos and the identification of relationships between the content of the videos and the content of the scientific articles mentioned. Consequently, they contribute to a comprehensive and contextualized analysis of scientific dissemination.

## 4 CONCLUSION

The Social4Science platform, as delineated in this study, facilitates the aggregation and examination of scientific data derived from social media, yielding significant insights concerning the propagation of scientific content. Through the analysis of these data, it is possible to identify trends, patterns, and knowledge gaps in discussions about scientific publications on social media. This instrument offers researchers and professionals in the scientific field crucial information, enabling them to adjust

communication strategies and promote scientific knowledge, thereby establishing a more effective connection with the general public. The data collected by the proposed platform can be used to establish significant correlations between different variables. These correlations provide a more profound understanding of the relationship between the popularity of a video on YouTube and the characteristics of the scientific article it references, taking into account factors such as research area, publication type, and country of origin. These analyses facilitate a comprehensive examination of the impact and repercussions of scientific publications on social media, thereby contributing to the understanding of the process of dissemination and the reach of scientific knowledge. The complete tool, developed with the source code of all the framework modules, will be made available in a GitHub repository for any community of interest.

## *Conflict of interest*

The authors declare no potential conflicts of interest.

## *Contribution statement*

**Conceptualization:** Thiago Magela Rodrigues Dias, Rafael Gonçalo Ribeiro, Patrícia Mascarenhas Dias
**Data curation:** Thiago Magela Rodrigues Dias, Rafael Gonçalo Ribeiro
**Formal Analysis:** Thiago Magela Rodrigues Dias, Rafael Gonçalo Ribeiro, Patrícia Mascarenhas Dias
**Methodology:** Thiago Magela Rodrigues Dias, Rafael Gonçalo Ribeiro, Patrícia Mascarenhas Dias
**Writing – Original Draft:** Rafael Gonçalo Ribeiro
**Writing – Review and Editing:** Thiago Magela Rodrigues Dias, Rafael Gonçalo Ribeiro, Patrícia Mascarenhas Dias

## *Statement of data consent*

All codes developed to build the platform can be noted in the following repository: *https://github.com/RafaelGoncalo/social4ScienceCLI*.

## REFERENCES


Araújo, R. F. (2020). Communities of attention networks: Introducing qualitative and conversational perspectives for altmetrics. *Scientometrics*, *124*(3), 1793–1809. *https://doi.org/10.1007/s11192-020-03566-7*

Da Fonseca, A. A., & Bueno, L. M. (2021). Breve panorama da divulgação científica brasileira no YouTube e nos podcasts. *Cadernos de Comunicação*, *25*(2). *https://doi.org/10.5902/2316882X63121*

Neto, J. R. S. (2018). Alcance da divulgação científica por meio do YouTube: estudo de caso no canal Meteoro Brasil. *Múltiplos Olhares em Ciência da Informação*, *8*(2).

Reale, M. V., & Martyniuk, V. L. (2016). Divulgação Científica no Youtube: a construção de sentido de pesquisadores nerds comunicando ciência. In *Congresso brasileiro de ciências da comunicação* (vol. 39, pp. 1–15).

CHAPTER 8

# DESIGN WITHOUT DATA? A STUDY OF METHODOLOGICAL TRANSPARENCY IN CONTEMPORARY DESIGN SCIENCE

**Jefferson Lewis Velasco**
*Pós-Design, Federal University of Santa Catarina, Brazil.*
**Email:** *jefferson.velasco@posgrad.ufsc.br*
ORCID: *https://orcid.org/0000-0002-1882-1785*

**Adilson Luiz Pinto**
*Department Information Science, Pós-Design,*
*Federal University of Santa Catarina, Brazil.*
ORCID: *https://orcid.org/0000-0002-4142-2061*

**Júlio Monteiro Teixeira**
*Pós-Design, Federal University of Santa Catarina, Brazil.*
ORCID: *https://orcid.org/0000-0002-9887-419X*

## ABSTRACT

This study investigated the methodological landscape of contemporary design science by analyzing 7,511 articles published across 10 leading journals in the field. The objective of this study was twofold: first, to ascertain the prevalence of qualitative, quantitative, and other forms of inquiry, and second, to reflect on the implications of methodological choices within design scholarship. The utilization of OpenAlex for the collection of metadata and ChatGPT-4o for the classification of abstracts based on

method-related keywords enabled the study to categorize articles as quantitative, qualitative, mixed methods, or inconclusive. The findings indicated that a mere 5.8% of the articles employed quantitative methods, while 14.28% utilized qualitative methods. Notably, 77.78% of the articles exhibited an absence of clear methodological signals, indicating a deficiency in methodological transparency. The application of topic modeling to inconclusive works revealed a preponderance of research that was conceptual, practice-based, or speculative in nature. These findings lent further credence to ongoing discourse regarding the dearth of methodological transparency and the underutilization of empirical strategies in design. The study's conclusion asserted that enhancing methodological articulation and establishing shared standards fortified the credibility and interdisciplinary recognition of design as a scientific field.

**KEYWORDS**: design research, research methods, bibliometrics, qualitative research, data-driven design, methodological transparency

**HOW TO CITE**: Lewis Velasco, J., Pinto, A. L., & Monteiro Teixeira, J. (2025). Design without data? A study of methodological transparency in contemporary design science. In A. Semeler (Ed.), Artificial Intelligence and Data Science Practices in Scientific Development, *Advanced Notes in Information Science, volume 8* (pp. 208-235). Pro-Metrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9331-4-5.116.

**COPYRIGHT**: © 2025 The author(s). This article is distributed under the terms of the CC BY-NC 4.0 license, which permits copying and redistribution of the material in any medium or format, adaptation, transformation, and building upon the material, provided that the license terms are followed.

## 1 INTRODUCTION

The production of scientific knowledge encompasses a broad spectrum of epistemological approaches, which are generally categorized based on the nature of their research methods. In general, methodologies are commonly divided into three categories: quantitative, qualitative, or mixed methods (Creswell & Creswell, 2018). Quantitative research is commonly linked to positivist

frameworks, which prioritize systematic methodologies, numerical data, and the attainment of objective, generalizable results. Conversely, qualitative research aligns with interpretivist traditions, emphasizing unstructured or semi-structured approaches, textual or visual data, and context-dependent interpretation. However, these distinctions can obscure the complex interdependencies between approaches, which, in practice, often overlap or are combined depending on disciplinary norms and research goals (Pilcher & Cortazzi, 2024). Contrary to the notion of these approaches representing diametrically opposed paradigms, they frequently operate in a complementary loop, reinforcing and enriching each other (Greene et al., 1989). Quantitative research frequently identifies general patterns or statistical regularities across extensive datasets, thereby offering a comprehensive understanding of phenomena and directing researchers toward domains that necessitate further investigation. Conversely, qualitative research has been demonstrated to excel at exploring the nuances of specific cases, uncovering contextual factors, subjective meanings, or anomalies that may remain invisible in aggregated data. Insights derived from qualitative inquiry frequently inform the formulation of novel hypotheses or the identification of variables to be tested quantitatively, thereby contributing to the continuous refinement of the research process (Tenny et al., 2025). As Pilcher and Cortazzi (2024) emphasize, this interdependence reflects the blurring of epistemological boundaries and underscores how real-world research often defies binary divisions (Figure 1).

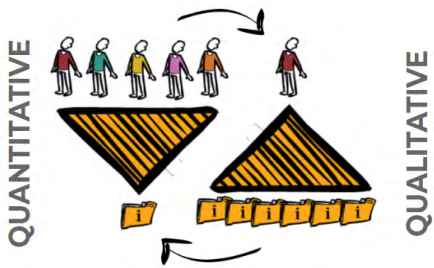


**Figure 1.** Complementary loop of quantitative and qualitative research. **Note**. Teixeira and Velasco (2024).

This dynamic relationship is particularly evident in fields such as design research, education, and human–computer interaction (HCI), where quantitative studies often measure performance indicators or behavioral patterns, while qualitative methods provide deeper insight into user experience and context (Van Turnhout et al., 2014). In the domain of design research, metrics such as user satisfaction or task efficiency across prototypes have been shown to reveal usability issues (Pinto et al., 2025). These issues often necessitate further study through methods such as interviews or observational techniques, which aim to enhance the comprehension of user responses (Weichbroth, 2019). In the field of education, standardized assessments have been shown to reveal disparities in learning outcomes across large populations. These disparities can be further elucidated through the use of classroom ethnographies, which offer a socio-emotional and cultural perspective on the underlying causes of these disparities (Mejeh et al., 2023). In the field of HCI, the utilization of analytics and A/B testing has emerged as a pivotal method for identifying interface issues. However, it is imperative to recognize the complementarity of think-aloud protocols and contextual inquiries, which unveil the underlying user behaviors and motivations. Across these domains, the interplay between data and interpretation—between breadth and depth—illustrates the evolving integration of research methods. A comprehensive study by Thelwall and Nevill (2021) found that qualitative research methods gained substantial prominence across academic disciplines between 1996 and 2019, signaling a shift toward broader acceptance of interpretive approaches. Notwithstanding the advent of big data and statistical modeling, qualitative methodologies—namely interviews, case studies, and ethnographies—have not only endured but have undergone an expansion in their scope. This tendency was particularly evident in the "social sciences" and "arts & humanities" fields, where qualitative inquiry has gained significant recognition and is actively promoted by journal editors, reviewers, and educators. Despite its continued status as a minority approach within certain scientific disciplines, qualitative research has firmly established itself as a mainstream component of academic scholarship.

Concomitantly, this growing acceptance has redirected attention to concerns regarding research quality and methodological rigor. In contradistinction to quantitative studies, which are

founded on standardized procedures and statistical verification, qualitative approaches are more difficult to reproduce, audit, or validate independently due to their interpretive and contextual nature (Cole et al., 2024; Harris et al., 2019). This inherent challenge in replication poses significant difficulties for the processes of peer review and academic assessment, particularly in instances where methodological procedures are either underreported or inconsistently applied. As Thelwall and Nevill (2021) observe, although qualitative methods—particularly interviews—are becoming more prevalent, their citation impact has diminished in numerous disciplines, potentially indicative of concerns regarding their reliability or scholarly value. To address these issues, journals have begun to adopt structured reporting frameworks, such as the COREQ checklist for interview and focus group studies (Tong et al., 2007), which promote greater transparency, rigor, and coherence in qualitative research practices. In light of these broader developments, it is imperative to investigate whether analogous dynamics are evident in the domain of design research. As a field historically grounded in creative practice, interpretive inquiry, and user-centered exploration, design shares many characteristics with disciplines that have embraced qualitative methodologies (Cross, 2001). However, despite the existence of anecdotal evidence and editorial preferences that appear to indicate a prevailing inclination toward qualitative approaches, there is a conspicuous absence of systematic data that would allow for the confirmation of this perception. If qualitative methods are indeed predominant, then design research may also be vulnerable to the same challenges related to transparency, reproducibility, and evaluative rigor. This concern is further compounded by the existence of adjacent modes of inquiry, such as speculative and critical design (SCD), which function beyond the confines of traditional empirical frameworks. These approaches are often grounded in critical theory and artistic practice, emphasizing conceptual provocation over data collection. This further complicates methodological classification and peer evaluation.

This study is an extension of this premise. It examines over 7,000 articles published in 10 prominent design science journals, and the objective is to determine whether the perceived preference for qualitative research is supported by empirical evidence. The objective of this study is to address the following research question: To what extent does the extant literature on design

science demonstrate a preference for qualitative methods over quantitative ones, and how consistent is this pattern across different journals? By mapping the methodological tendencies in contemporary design research, the study contributes to a clearer understanding of the field's knowledge production practices and highlights opportunities for increased methodological balance and transparency.

## 2 THEORY

### *2.1 Research methodologies: Quantitative, qualitative, mixed, and alternative approaches*

Academic research is commonly structured around three methodological paradigms: quantitative, qualitative, and mixed methods. Quantitative research is rooted in positivist or post-positivist traditions, emphasizing measurement, numerical analysis, and statistical inference to test hypotheses or identify patterns across populations (Babbie, 2016; Creswell & Creswell, 2018). Conversely, qualitative research is predicated on interpretivist or constructivist worldviews, with the objective being to comprehend meanings, behaviors, and experiences through in-depth, context-sensitive approaches such as interviews, observations, and document analysis (Denzin & Lincoln, 2011). Mixed methods research intentionally integrates both paradigms, leveraging the strengths of each to provide a more comprehensive understanding of a research problem (Creswell & Clark, 2018). While these categories are often presented as distinct, in practice, they frequently overlap, reflecting the complexity of real-world inquiry and the increasing recognition of methodological pluralism. In addition to these empirical approaches, some research—particularly in fields such as design, philosophy, and the arts—adopts nonempirical or practice-based formats. These include theoretical or conceptual studies, which aim to develop or critique ideas rather than collect data, and practice-based research, in which creative activity itself becomes a method of inquiry and a source of knowledge (Barrett & Bolt, 2010; Frayling, 1993). Despite their deviation from conventional empirical paradigms, speculative

design, design fiction, and other reflective or critical approaches also play important roles in knowledge production. These inquiries contribute meaningfully to academic discourse by expanding the epistemological boundaries of research.

A considerable number of scholars posit that theoretical, conceptual, and practice-based studies fall within the ambit of qualitative research. However, there is a counterargument positing that theoretical, conceptual, and practice-based research should be regarded as distinct modes of inquiry rather than as subcategories of qualitative research. While these approaches may be considered similar in terms of their interpretive orientation and their lack of reliance on numerical data, they differ in the assumptions they operate under and the types of knowledge they produce. For instance, practice-based research is frequently founded on the process of creative production and the generation of insights through the act of making, as opposed to observation or interaction with participants. As Biggs and Büchler (2007) emphasize, this form of inquiry is epistemologically unique and should not be assessed by the same criteria applied to traditional qualitative or quantitative studies. It is imperative to approach all nonnumerical research as qualitative risks. This approach entails a comprehensive examination of the fundamental differences in research logic and goals, which are often oversimplified in other frameworks. A fundamental distinction can be identified in the absence of empirical data collection or participant involvement, which are hallmarks of most qualitative methodologies. Qualitative research, on the other hand, typically involves the use of interviews, ethnography, or document analysis to understand social phenomena. In contrast, theoretical and conceptual inquiries rely on argumentation, synthesis, or critique without gathering first-hand data. Conversely, practice-based research may entail self-reflection, autoethnography, or artifact generation, without the involvement of external subjects or replicable datasets. As posited by Biggs and Büchler (2007), practice-based research "encompasses a creative output as an integral component of the research process," thereby situating it beyond the empirical framework of both qualitative and quantitative paradigms.

The epistemological foundations of these approaches further substantiate their distinctiveness. Qualitative research is generally situated within interpretivist paradigms, which seek

to understand meaning from the perspective of human actors. Conversely, theoretical research frequently draws from rationalist or critical traditions, utilizing logic, conceptual analysis, and dialectics as primary methodologies. Practice-based research, as articulated by Borgdorff (2012), is predicated on constructivist and artistic epistemologies, wherein knowledge is interwoven with and manifests through praxis. This diversity of foundations underscores the methodological heterogeneity of nonquantitative inquiry and highlights the limitations of treating them as interchangeable under a single qualitative label. Despite these differences, it is evident that theoretical, conceptual, and practice-based research do not align with quantitative paradigms, which are rooted in measurement, hypothesis testing, and statistical inference. Consequently, while they are distinct from both quantitative and traditional qualitative research, they may nevertheless be subject to similar challenges that confront qualitative methods. These include concerns about reproducibility, transparency, and methodological rigor—particularly when studies lack formal frameworks or clear documentation of procedures. The recognition of these common challenges underscores the necessity for customized evaluation standards that acknowledge the distinct contributions and limitations of each research modality. To better illustrate this discussion, Table 1 is presented to clarify the differences and similarities of the discussed methods.

**Table 1.** Differences and similarities in methodological characteristics. **Note**. Prepared by authors.

| Feature | Quantitative research | Qualitative research | Theoretical/ conceptual research | Practice-based research |
|---|---|---|---|---|
| Data collection | Yes (structured and numerical) | Yes (textual, visual, and contextual) | No | Sometimes (via reflection or logs) |

| Feature | Quantitative research | Qualitative research | Theoretical/ conceptual research | Practice-based research |
|---|---|---|---|---|
| Human participants | Yes (e.g., surveys and experiments) | Yes (e.g., interviews and observations) | Rare | Rare |
| Nature of data | Numerical | Textual, visual, and verbal | Argument-based | Creative work + reflective text |
| Main output | Statistical findings | Thematic or narrative findings | Theories, frameworks, and critiques | Artifacts + reflection/insight |
| Epistemological roots | Positivism/ post-positivism | Interpretivism | Rationalism and constructivism | Constructivism, aesthetics, and critical |
| Evaluation criteria | Validity, reliability, and replicability | Credibility, transferability, and trustworthiness | Coherence, logic, and originality | Reflexivity and process transparency |

## *2.2 Limited use of data-driven approaches in design research*

Despite the growing emphasis on methodological rigor across disciplines, design research continues to demonstrate a relative scarcity of data-reliant studies, particularly those grounded in quantitative or systematically collected empirical evidence. A significant proportion of the field's scholarly output may persist in being anchored in interpretive, conceptual, or practice-based methodologies. This could be indicative of a predilection for exploration, reflection, and the construction of meaning, as opposed to the testing of hypotheses or the generalization of statistical findings. While this orientation reflects the creative and user-centered foundations of design, it also limits the adoption of methodologies that enable broader pattern identification, replicability,

and comparability across studies (Escudero-Mancebo et al., 2023). One illustrative example of this tendency is SCD. Introduced by Dunne and Raby (2013), SCD repositions design as a discursive practice aimed at questioning prevailing technological and cultural assumptions rather than solving practical problems. This approach emerged as a response to the instrumentalism characteristic of mainstream design practice, proposing instead that design should serve as a tool for reflection, critique, and cultural commentary. The methods employed by this group are primarily conceptual in nature. Designers create fictional scenarios or artifacts with the intention of provoking debate, raising ethical concerns, or reframing societal issues. Consequently, SCD functions beyond the confines of conventional empirical frameworks, eschewing formal data collection and seldom engaging directly with users or environments (Johannessen et al., 2019).

While SCD has expanded the epistemological boundaries of design by legitimizing critique, provocation, and conceptual exploration, its ambiguous methodological status also exposes a broader vulnerability within the field. Design research frequently functions in the absence of a definitive consensus regarding evidentiary standards or methodological rigor, a phenomenon that is particularly evident in studies that do not draw upon empirical data. The multifaceted nature of this phenomenon bestows researchers with methodological flexibility. However, this flexibility can also result in a lack of orientation and observed "lack of rigor," an issue that has been noted in empirical design studies (Toh et al., 2014). Consequently, projects that eschew formal data collection, whether speculative, conceptual, or artistic, may encounter challenges in communicating their contributions in ways that are auditable, reproducible, or broadly comparable (Timperley et al., 2021). This ambiguity complicates peer review, editorial evaluation, and scholarly dialogue, especially in interdisciplinary settings where expectations around transparency, validity, and impact are shaped by more established research paradigms. Concurrently, design is undergoing a substantial transformation as it increasingly interfaces with technological domains such as UX, CX, service design, and digital product development. Although interaction data are more accessible than ever, they remain underutilized in many traditional design workflows due to methodological misalignment and integration barriers. As Quiñones-Gómez et al. (2025) observe, the integration of

data-driven insights into established design paradigms remains a complex and under-explored area, underscoring the necessity for coherent frameworks that facilitate the integration of data and design. These domains underscore interaction data, behavioral patterns, and performance metrics (Pinto et al., 2025), components that inherently favor data-driven inquiry. As the availability of data increases through digital platforms (Hilbert & López, 2011), there is an increasing expectation for designers and researchers to adopt empirical methods capable of capturing and interpreting this information meaningfully. In light of these arguments, the subsequent section will examine this emerging tension by analyzing the opportunities and responsibilities that accompany data availability in design research and practice.

### *2.3 Design in a data-producing society*

The proliferation of digital technologies and interconnected systems has led to an era where data are constantly generated, captured, and stored, thereby transforming the very fabric of modern life (Hilbert & López, 2011). This transition toward a society that produces data offers novel opportunities for understanding user behavior, system performance, and social dynamics—opportunities that remain largely unexplored in conventional design research. As Oppermann and Munzner (2020) suggest, "data-first design studies" reverse the standard model by allowing real-world data to drive design insights and decisions, rather than starting with design questions or assumptions. Design, particularly in its digital and service-oriented manifestations, has become profoundly intertwined with data ecosystems (Velasco et al., 2025). As products evolve into platforms and services transition to digital channels, designers now have access to near real-time, granular, and scalable feedback. Interaction logs, performance metrics, and analytics tools are increasingly being used to guide design decisions. According to Quiñones-Gómez et al. (2025), "data-driven design is a methodology that relies on quantitative and qualitative data to inform and shape design decisions in digital product development," thereby highlighting this emergent shift in practice. This may encompass a wide range of data, including clickstream data, A/B test results, heatmaps, telemetry, and usage logs. Each of these data sources offers

valuable insights into how users interact with designed systems. These data sources serve to complement qualitative methods and provide evidence that can validate design decisions, reveal usage patterns, and identify opportunities for improvement that may elude purely interpretive approaches. For instance, Ebel et al. (2023) demonstrate how automotive interface telemetry, when visualized and analyzed, can directly inform UX design and drive iterative product refinement.

Notwithstanding this potential, the systematic use of data in design research and practice may remain limited. A considerable number of design projects continue to prioritize experiential and conceptual outputs, while neglecting to consider the potential of behavioral data to inform or evaluate outcomes. Walny et al. (2020) describe how, in data visualization design, even when data are central, design focus often remains on artifact presentation and encoding decisions rather than on structured, behavioral data analysis. The observed discrepancy is indicative of not only epistemological traditions but also a dearth of methodological frameworks and a paucity of literacy in data-driven techniques among design professionals. As design increasingly intersects with areas such as UX, CX, and digital product development—domains where analytics and experimentation are routine—the need for data fluency becomes more relevant (Ebel et al., 2023). This evolving context necessitates a reexamination of the methods by which evidence is defined, gathered, and interpreted in design. As design becomes increasingly intertwined with data-rich environments, it is imperative to understand the methodological foundations of the field. Prior to advocating for greater integration of data-reliant or quantitative approaches, it is imperative to investigate the current state of research practices within the discipline. To provide a foundation for this reflection, the prevalence of methodological paradigms must be mapped, including those of a qualitative, quantitative, or nonempirical nature. By first identifying how design research is currently conducted, the field can meaningfully engage with questions of methodological rigor, evidentiary standards, and the role of data in shaping design knowledge. In this context, the integration of quantitative and computational methods into design represents more than a mere technical evolution; it is, in essence, a contextual response to the epistemic and societal conditions that are characteristic of the digital age.

## 3 METHODOLOGY

This study employs a data-driven approach to map the methodological orientation of contemporary design research. The investigation commenced with the selection of 10 prominent, active journals in the field of design science. These journals were chosen for their relevance and academic impact, as indicated by metrics such as CiteScore, SCImago Journal Rank (SJR), and impact factor, as presented in Table 2. A comprehensive dataset was compiled on May 30, 2025, using the OpenAlex database as a source. This dataset contains metadata from all articles published in the aforementioned journals, resulting in a total sample of 7,511 works. Subsequently, the abstracts of each article were analyzed using ChatGPT-4o to ascertain the presence of keywords indicative of either qualitative or quantitative research methodologies. Articles that lacked sufficient information for classification were labeled as inconclusive, with the understanding that they may represent theoretical, conceptual, or practice-based studies.

**Table 2.** Sources, relevance, and impact. **Note**. Prepared by authors.

| Journal | Year first published | Scope description | CiteScore | SJR | Impact factor |
|---|---|---|---|---|---|
| *Design Studies* | 1979 | It focuses on developing an understanding of design processes across various domains, including engineering, product design, architectural and urban design, and systems design. | 6.7 | 1.231 | 3.2 |

| Journal | Year first published | Scope description | CiteScore | SJR | Impact factor |
|---|---|---|---|---|---|
| *The Design Journal* | 1998 | It covers all aspects of design, providing a forum for design scholars, professionals, educators, and managers worldwide. | 1.4 | 398 | 0.8 |
| *Journal of Design History* | 1988 | It embraces the history of a range of design-related subjects, from furniture to product design, graphic design, craft, fashion, textiles, architectural interiors, and exhibitions. | 0.8 | 166 | 0.3 |
| *International Journal of Design* | 2007 | A peer-reviewed, open-access journal devoted to publishing research papers in all fields of design, including industrial design, visual communication design, interface design, and more. | 4.5 | 876 | 1.6 |

| Journal | Year first published | Scope description | CiteScore | SJR | Impact factor |
|---|---|---|---|---|---|
| *Design Issues* | 1984 | The first American academic journal to examine design history, theory, and criticism, provoking inquiry into cultural and intellectual issues surrounding design. | 1.3 | 0.24 | 0.4 |
| *Journal of Engineering Design* | 1990 | It provides a forum for the publication of high-quality, peer-reviewed papers on engineering design, covering design theory, methodology, and practice. | 5.2 | 603 | 2.5 |
| *CoDesign* | 2005 | It focuses on collaborative and participatory design processes across a range of disciplines, including design, arts, and social sciences. | 6.1 | 1.085 | 2.0 |

| Journal | Year first published | Scope description | CiteScore | SJR | Impact factor |
|---|---|---|---|---|---|
| *Design and Culture* | 2009 | It explores the cultural significance of design and its impact on society, combining perspectives from design studies, cultural studies, and related fields. | 1.8 | 278 | 0.7 |
| *Design Science* | 2015 | It publishes interdisciplinary research on all aspects of design science, including theory, methodology, and practical applications in engineering, architecture, computing, and other design fields. | 5.7 | 662 | 2.82 |
| *International Journal of Design Creativity and Innovation* | 2013 | It explores creativity and innovation in design, emphasizing multidisciplinary and interdisciplinary approaches to creative processes. | 3.1 | 452 | 1.2 |

### 3.1 Data collection

The dataset examined in this study was retrieved from OpenAlex and comprised metadata for 7,511 academic publications in the field of design science. After the removal of records lacking an abstract from the dataset, a total of 2,052 documents were obtained, constituting the working corpus. The "abstract" field was selected as the primary source for analysis, under the assumption that it would contain methodological information relevant to classifying the research approach adopted in each paper. Four abstracts were excluded from topic modeling due to malformed or corrupted content that failed to yield any usable features for analysis.

### 3.2 Data analysis

To identify the methodological approach employed by each paper ("quantitative," "qualitative," or "inconclusive"), a rule-based classification method was applied to the text of the abstracts. The employment of regular expression pattern matching was instrumental in the identification of keywords commonly associated with quantitative or qualitative research methodologies. In the event that an abstract contained indicators from both categories, it was labeled "both." In the event that no such findings were present, the result was designated as "inconclusive." This approach was selected to facilitate rapid, large-scale screening without the need for manual annotation, a process that was further expedited by the implementation of artificial intelligence (AI).

- **Keywords used to identify quantitative methods:** Survey, regression, statistical analysis, quantitative, experiment, data set, dataset, quantitatively, questionnaire, correlation, ANOVA, t-test, descriptive statistics, and sample size.

- **Keywords used to identify qualitative methods:** interview, focus group, ethnography, case study, qualitative, observation, thematic analysis, content analysis, narrative, grounded theory, field notes, and participant observation.

To further understand the content of the abstracts labeled as "inconclusive" (n = 1,596), topic modeling was applied with AI assistance using non-negative matrix factorization (NMF). The abstracts were initially converted into a term-document matrix utilizing TF-IDF vectorization, with the top 1,000 terms identified as the most informative, and stop words in English removed. The NMF algorithm was implemented with five components, which corresponded to five latent topics. The abstracts were then assigned to a topic based on the component with the highest weight. The top 10 keywords per topic were extracted to support the interpretation and labeling of topics. This analysis successfully described the majority of the inconclusive sample.

## 4 RESULTS

This section presents the findings derived from the classification and analysis of 2,052 articles published in 10 leading design science journals. The initial classification revealed that only a small fraction of works employed quantitative (5.8%) or qualitative (14.28%) methods, while the majority (77.78%) could not be confidently categorized (Table 3).

**Table 3.** Work classification. **Note**. Prepared by authors.

| Classification | Frequency (%) |
|---|---|
| Quantitative | 5.8 |
| Qualitative | 14.28 |
| Mixed | 2.14 |
| Inconclusive | 77.78 |

To further examine the nature of these inconclusive works, topic modeling was applied to their abstracts, uncovering five dominant thematic clusters that illustrate the methodological diversity—and ambiguity—within contemporary design research. The

results are structured in two parts: (1) the frequency and distribution of methodological classifications, and (2) a qualitative interpretation of themes emerging from the inconclusive subset (Table 4).

**Table 4.** Inconclusive sample analysis. **Note**. Prepared by authors.

| Topic | Frequency (%) | Top terms in topic | Summary |
|---|---|---|---|
| Design research and methodology | 26.25% | Design, research, practice, process, knowledge, thinking, education, paper, framework, methods | These papers discuss design as a research discipline, often referencing conceptual or pedagogical frameworks without specifying methods. |
| Conference/ event metadata | 6.52% | 2019, scissors, pp, dundee, running, 13th, bletcher, valentine, cruickshank | This topic includes event references, likely representing metadata from conference proceedings rather than substantive content. |
| Web/ indexing artifacts | 5.83% | Search, doi, icon, author, university, org, https, issues, institute, site | These records are probably noise—scraped metadata, broken abstracts, or entries containing only web or reference boilerplate. |
| Product design and engineering | 24.12% | Product, products, development, method, process, engineering, based, model, use, user | These abstracts discuss technical aspects of product or system design, possibly in engineering contexts, but without mentioning how the research was conducted. |

| Topic | Frequency (%) | Top terms in topic | Summary |
|---|---|---|---|
| Digital and social innovation | 37.03% | Social, new, people, paper, cultural, innovation, digital, public, service, objects | This theme focuses on digital transformation, cultural change, or public service innovation, often theoretical or reflective in tone. |
| Topic not identified | 0.25% | N/A | No topic could be identified in these papers. |

Table 3 presents a summary of the distribution of methodological classifications across the sample. A total of 119 articles (5.8%) were identified as quantitative, while 293 articles (14.28%) were classified as qualitative. The majority of articles (1,596, or 77.78%) were classified as inconclusive, indicating an absence of clear references to methodological frameworks typically associated with empirical studies. This distribution indicates that, while empirical research is present in design science, it is not yet the predominant approach. The preponderance of inconclusive articles lends credence to the notion that a significant portion of the field's research remains anchored in interpretive, conceptual, or practice-based methodologies, which do not depend on explicit methodological indicators discernible through keyword analysis. In light of the inconclusive findings from Table 4, it is evident that the predominant cluster pertains to the domain of design education, with a particular emphasis on pedagogical methodologies, the attainment of learning outcomes, and the development of curricula. These subjects frequently prioritize the cultivation of reflective and experiential knowledge over formal empirical validation. The second most prominent theme involves sustainability and social innovation, areas that are often explored through speculative or value-driven approaches that defy easy classification. Other clusters include design theory and methodology, user-centered processes, and emerging technologies. These other clusters may involve conceptual work or practice-based inquiry without explicit methodological articulation. Conference/event metadata and web/indexing artifacts were identified as likely

noise and deemed irrelevant for the purposes of this research, as they do not contribute to the methodological orientation or thematic content of the articles. Furthermore, four papers in the dataset could not be reliably assigned to any thematic cluster, suggesting insufficient or ambiguous abstract content for topic modeling.

## 5 DISCUSSION

The findings presented in this study offer a comprehensive overview of the methodological landscape in contemporary design science literature. Of the 2,052 articles that were subjected to analysis, a negligible proportion were classified as quantitative (5.8%) or qualitative (14.28%), with a minimal number employing mixed methods (2.14%). Most notably, the majority (77.78%) were classified as "inconclusive," exhibiting a lack of clear methodological markers traditionally associated with empirical studies. This finding resonates with persistent concerns articulated within the domain of design theory, particularly concerning the epistemological foundations of the field and the frequently ambiguous nature of its knowledge production practices (Cross, 2001). This methodological opacity appears to confirm the dominance of interpretive, conceptual, or practice-based traditions within design research—traditions that frequently resist classification using empirical criteria. As Pilcher and Cortazzi (2024) contend, design scholarship functions at the nexus of numerous epistemological paradigms, where the distinctions between empirical, speculative, and artistic modes of inquiry are permeable. However, the limited availability of empirical transparency presents significant challenges, particularly in light of the mounting calls for methodological rigor and auditability across various disciplines (Cole et al., 2024; Harris et al., 2019). The observed discrepancy between qualitative and quantitative studies within the identifiable subset—where qualitative works appear almost three times more common—further reinforces the perception that design scholarship tends to privilege interpretive over generalizable analysis. This phenomenon, however, does not inherently pose any significant challenges. In fact, it can be viewed as a reflection of the historical emphasis that design has placed on user-centered, contextual, and reflexive knowledge creation.

However, as Van Turnhout et al. (2014) observe, this orientation may inadequately prepare the field for engagement with evidence standards and evaluative frameworks that are increasingly dominant in adjacent domains, such as HCI and service design.

The application of topic modeling to the 1,596 inconclusive articles provides further insights. The largest cluster, "digital and social innovation" (37.03%), corresponds to domains that are typically associated with reflective, ethical, and societal concerns. These works generally address emergent challenges using speculative, conceptual, or value-driven perspectives, consistent with the principles of SCD (Dunne & Raby, 2013). While such contributions are valuable, they may not meet traditional academic standards of evidence and reproducibility. The second most prevalent category, "design research and methodology" (26.25%), encompasses works that delve into the foundational principles, conceptual frameworks, and pedagogical dimensions of design as a discipline. These papers frequently engage with abstract or philosophical discussions about design thinking and research practice, but they do so without specifying data sources or procedural details. This further reinforces the prevalence of conceptual or exploratory work in the field. The third cluster, "product design and engineering" (24.12%), demonstrates engagement with technical systems and user-centered tools, yet exhibits minimal methodological transparency. This phenomenon may be indicative of a practice-based reporting style, which prioritizes the presentation of evidence over the exposition of underlying principles. Alternatively, it could be attributed to the influence of engineering disciplines, where methodological descriptions are implicit but not explicitly articulated.

## 6 CONCLUSION

This study reveals that design science embraces a diverse array of knowledge-making strategies, many of which diverge from conventional empirical norms. As design increasingly interfaces with data-rich domains such as UX, CX, and digital product development, this lack of methodological articulation may hinder its ability to communicate contributions effectively within broader scientific discourses. Furthermore, the dearth of shared evidentiary standards jeopardizes the marginalization of entire

subfields—such as speculative or conceptual design—whose value is arduous to assess using conventional academic metrics. The findings indicate that a considerable proportion of design research either evades or exhibits an absence of the methodological transparency that is generally anticipated in other disciplines. While reflective, speculative, and conceptual approaches are integral to the field, their growing prevalence underscores the need for more precise criteria to distinguish between modes of inquiry and evaluate their scholarly merit. In the absence of a more precise methodological articulation, there is a risk that design research may be misclassified, misunderstood, or undervalued, particularly in fields where empirical grounding is widely regarded as the gold standard of credibility. This methodological opacity is indicative of a discipline that is deeply rooted in exploration, practice, and reflection—forms of inquiry that resist facile classification and rarely conform to the reproducibility and auditability standards of the natural and social sciences. As design becomes increasingly intertwined with technology, reliance on systematic evaluation and data fluency grows. Consequently, the credibility and relevance of design will be contingent on the development of stronger methodological clarity and accountability. In this context, this study serves as a preliminary step toward elucidating the methodological composition of design research. By mapping the distribution of empirical and nonempirical approaches, the study contributes to ongoing efforts to rethink what constitutes valid evidence in design—and how diverse modes of inquiry can be recognized, validated, and integrated into a more inclusive and methodologically reflective research culture.

Notwithstanding its contributions, this study is not without limitations. First, the classification system was dependent on automated keyword analysis in abstracts. While this method is scalable and efficient, it has the potential to overlook methodological nuances or frameworks that are discussed exclusively in full texts. Second, although topic modeling offers insight into the inconclusive subset, it remains an interpretive tool, subject to subjective interpretation. Third, the keyword sets utilized may not fully encompass the range of terms associated with qualitative or quantitative research, potentially leading to underrepresentation. Furthermore, the expansion of the dataset to encompass a more extensive array of works and a wider spectrum of publication types would enhance the generalizability of the findings. A

further structural limitation is evident in the strategy employed for journal selection. In the absence of a formal taxonomy of design science, the focus on 10 prominent journals—though methodologically justifiable—may introduce epistemological bias. These publications may include a disproportionate representation of particular subfields or methodological preferences, thereby constricting the breadth of the analysis. This may reveal a more extensive issue: the field could benefit from the development of a widely accepted taxonomy that defines its epistemic boundaries, paradigms, and methodological standards. The implementation of such a framework has the potential to enhance clarity, facilitate comparative research, and establish comprehensive evaluation criteria across the discipline. In essence, the mapping of methodological tendencies presented herein establishes a foundational framework for subsequent investigations into the epistemological dynamics of design research. Subsequent studies could build on this work by refining classification methods (e.g., through manual coding or supervised machine learning), increasing the scope of analysis, and exploring correlations between method and research impact. It is imperative to enhance methodological transparency and cultivate a unified lexicon of inquiry to ensure the advancement of the rigor, relevance, and recognition of design as a scientific discipline.

## *Conflict of interest*

The authors declare that there are no conflicts of interest.

## *Contribution statement*

**Conceptualization, Methodology, Formal Analysis, Investigation, Writing Original Draft:** Jefferson Lewis Velasco.
**Supervision:** Júlio Monteiro Teixeira.
**Writing Review:** Adilson Luiz Pinto, Júlio Monteiro Teixeira.

### *Statement of data consent*

The datasets generated during the development of this study have been deposited in Google Drive and are accessible at: *https://drive.google.com/drive/folders/1lOn8ZsrT8uvDzOcIxqmEB-cqjHHlo1UoU?usp=sharing*.

## REFERENCES


Babbie, E. (2016). ***The practice of social research***. Cengage Learning.

Barrett, E., & Bolt, B. (Eds.). (2010). ***Practice as research: Approaches to creative arts enquiry***. I.B. Tauris.

Biggs, M. A. R., & Büchler, D. (2007). Rigor and practice-based research. ***Design Issues***, *23*(3), 62–69. *https://doi.org/10.1162/desi.2007.23.3.62*

Borgdorff, H. (2012). ***The conflict of the faculties: Perspectives on artistic research and academia***. Leiden University Press.

Cole, N. L., Ulpts, S., Bochynska, A., Kormann, E., Good, M., Leitner, B., & Ross-Hellauer, T. (2024). ***Reproducibility and replicability of qualitative research: An integrative review of concepts, barriers and enablers***. Center for Open Science. *https://doi.org/10.31222/osf.io/n5zkw*

Creswell, J. W., & Clark, V. L. P. (2018). ***Designing and conducting mixed methods research***. SAGE Publications, Inc.

Creswell, J. W., & Creswell, J. D. (2018). ***Research design: Qualitative, quantitative, and mixed methods approaches***. SAGE Publications, Inc.

Cross, N. (2001). Designerly ways of knowing: Design discipline versus design science. ***Design Issues***, *17*(3), 49–55. *https://doi.org/10.1162/074793601750357196*

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2011). ***The SAGE handbook of qualitative research***. SAGE Publications, Inc.

Dunne, A., & Raby, F. (2013). Speculative everything: Design, fiction, and social dreaming. The MIT Press.

Ebel, P., Gülle, K. J., Lingenfelder, C., & Vogelsang, A. (2023). Exploring millions of user interactions with ICEBOAT: Big data analytics for automotive user interfaces. In *AutomotiveUI '23: Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 81–92). *https://doi.org/10.1145/3580585.3607158*

Escudero-Mancebo, D., Fernández-Villalobos, N., Martín-Llorente, Ó., & Martínez-Monés, A. (2023). Research methods in engineering design: A synthesis of recent studies using a systematic literature review. *Research in Engineering Design*, *34*(2), 221–256. *https://doi.org/10.1007/s00163-022-00406-y*

Frayling, C. (1993). Research in art and design. *Royal College of Art Research Papers*, *1*(1). *https://researchonline.rca.ac.uk/384/*

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255–274. *https://doi.org/10.3102/01623737011003255*

Harris, J. K., Combs, T. B., Johnson, K. J., Carothers, B. J., Luke, D. A., & Wang, X. (2019). Three changes public health scientists can make to help build a culture of reproducible research. *Public Health Reports*, *134*(2), 109–111. *https://doi.org/10.1177/0033354918821076*

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, *332*(6025), 60–65. *https://doi.org/10.1126/science.1200970*

Johannessen, L. K., Keitsch, M. M., & Pettersen, I. N. (2019). Speculative and critical design—Features, methods, and practices. *Proceedings of the Design Society: International Conference on Engineering Design*, *1*(1), 1623–1632. *https://doi.org/10.1017/dsi.2019.168*

Mejeh, M., Hagenauer, G., & Gläser-Zikuda, M. (2023). Mixed methods research on learning and instruction—Meeting the challenges of multiple perspectives and levels within a complex field. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, *24*(1), Article 1. *https://doi.org/10.17169/fqs-24.1.3989*

Oppermann, M., & Munzner, T. (2020). Data-first visualization design studies. In *2020 IEEE Workshop on Evaluation and Beyond—Methodological Approaches to Visualization (BELIV)* (pp. 74–80). *https://doi.org/10.1109/beliv51497.2020.00016*

Pilcher, N., & Cortazzi, M. (2024). "Qualitative" and "quantitative" methods and approaches across subject fields: Implications for research values, assumptions, and practices. *Quality & Quantity*, *58*(3), 2357–2387. *https://doi.org/10.1007/s11135-023-01734-4*

Pinto, A. L., Teixeira, J. M., & Velasco, J. L. (2025). Understanding design metrics: A theoretical model for application and evaluation. *AWARI*, *6*, 1–10. *https://doi.org/10.47909/awari.833*

Quiñones-Gómez, J. C., Mor, E., & Chacón, J. (2025). Data-driven design in the design process: A systematic literature review on challenges and opportunities. *International Journal of Human–Computer Interaction*, *41*(4), 2227–2252. *https://doi.org/10.1080/10447318.2024.2318060*

Teixeira, J. M., & Velasco, J. L. (2024). *Design para Negócios Digitais: Qualitativo x Quantitativo* [Google Presentation slides]. Undergraduate Program of Design—University of Santa Catarina. *https://docs.google.com/presentation/d/1ESCASgoe01qYqiWogJmxeqceMv-e1JuHcsxc_uAB2WM/edit?usp=sharing*

Tenny, S., Brannan, J. M., & Brannan, G. D. (2025). Qualitative study. In *StatPearls*. StatPearls Publishing. *http://www.ncbi.nlm.nih.gov/books/NBK470395/*

Thelwall, M., & Nevill, T. (2021). Is research with qualitative data more prevalent and impactful now? Interviews, case studies, focus groups and ethnographies. *Library & Information Science Research*, *43*(2), Article 101094. *https://doi.org/10.1016/j.lisr.2021.101094*

Timperley, C. S., Herckis, L., Goues, C. L., & Hilton, M. (2021). Understanding and improving artifact sharing in software engineering research. *Empirical Software Engineering*, *26*(4). *https://doi.org/10.1007/s10664-021-09973-5*

Toh, W. X., Hashemi Farzaneh, H., Kaiser, M. K., & Lindemann, U. (2014). Analysis of empirical studies in design research. In *DS 77: Proceedings of the DESIGN 2014 13th International Design Conference* (pp. 59–70).

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, *19*(6), 349–357. *https://doi.org/10.1093/intqhc/mzm042*

Van Turnhout, K., Bennis, A., Craenmehr, S., Holwerda, R., Jacobs, M., Niels, R., Zaad, L., Hoppenbrouwers, S., Lenior, D., & Bakker, R. (2014). *Design patterns for mixed-method research in HCI*. In *NordiCHI '14. Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. *https://doi.org/10.13140/2.1.4701.2643*

Velasco, J. L., Teixeira, J. M., & Pinto, A. L. (2025). Validating personas for better communication: A structured model for low-resource contexts. *Iberoamerican Journal of Science Measurement and Communication*, *5*(3), Article 3. *https://doi.org/10.47909/ijsmc.236*

Walny, J., Frisson, C., West, M., Kosminsky, D., Knudsen, S., Carpendale, S., & Willett, W. (2020). Data changes everything: Challenges and opportunities in data visualization design handoff. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 12–22. *https://doi.org/10.1109/tvcg.2019.2934538*

Weichbroth, P. (2019). A mixed-methods measurement and evaluation methodology for mobile application usability studies. *Annals of Computer Science and Information Systems*, *20*, 101–106. *https://doi.org/10.15439/2019F299*

CHAPTER 9

# BIBLIOGRAPHIC ANALYSIS OF SCIENTIFIC LITERATURE ON HEALTH KNOWLEDGE MANAGEMENT

**Hossein Ghalavand**
*Department of Medical Library and Information Science, Abadan University of Medical Sciences, Iran.*
ORCID: *https://orcid.org/0000-0001-7647-5449*

**Reza Varmazyar**
*Department of Documentation, Carlos III University of Madrid, Spain.*
ORCID: *https://orcid.org/0000-0001-6013-261X*

**Saied Shirshahi**
*Department of Medical Library and Information Sciences, School of Management and Medical Information Sciences, Isfahan University of Medical Sciences, Iran.*
**Email:** *saeid.shirshahi@gmail.com*
ORCID: *https://orcid.org/0000-0001-9039-3650*

## ABSTRACT

This study utilized a scientometric approach to examine the landscape of health knowledge management (HKM), employing co-occurrence analysis to map thematic developments from 1990 to 2023. The data were retrieved from Web of Science and PubMed, and subsequently analyzed using VOSviewer and Excel 2019 to identify key patterns and collaborations. The findings indicated that the United States was a leader in research output, with Wright A and Sittig DF being particularly influential contributors. The co-occurrence map was comprised of four primary clusters:

(1) the impact of the Coronavirus 2019 (COVID-19) on health information management, (2) strategies to enhance healthcare performance and systems, (3) electronic medical records and related challenges, and (4) advancements in big data, Internet technologies, and foundational research. These clusters underscored pivotal domains that influenced HKM, accentuating the significance of effective knowledge management methodologies, particularly in the context of technological and global health transformations. The study emphasized the necessity of strategic resource allocation, investment in technological infrastructure, and international collaboration to enhance healthcare outcomes through effective knowledge dissemination and management.

**KEYWORDS**: co-occurrence, health knowledge management, bibliographic analysis, data science, research trends, data visualization

**HOW TO CITE**: Ghalavand, H., Varmazyar, R., & Shirshahi, S. (2025). Bibliographic analysis of scientific literature on health knowledge management. In A. Semeler (Ed.), Artificial Intelligence and Data Science Practices in Scientific Development, *Advanced Notes in Information Science, volume 8* (pp. 236-256). Pro-Metrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9331-4-5.117.

**COPYRIGHT**: © 2025 The author(s). This article is distributed under the terms of the CC BY-NC 4.0 license, which permits copying and redistribution of the material in any medium or format, adaptation, transformation, and building upon the material, provided that the license terms are followed.

## 1 INTRODUCTION

In recent years, knowledge management has gained significant attention across various sectors and organizational sizes. Its presence has been observed in small niche businesses, international firms, research institutes, and universities. Hansen and his colleagues, however, posit that knowledge management is not a recent development. The most prominent illustrations of this phenomenon include the proprietors of family-run businesses who have transferred their commercial acumen to their offspring, master craftsmen who have meticulously imparted

their trades to apprentices, and workers who have engaged in the exchange of ideas and expertise within the workplace (Hansen et al., 2005). It is, in essence, a substantial expanse of "comprising creation, acquisition, collation, sharing, use, reuse, and capitalization of knowledge in an organization" (Pandey, 2016, p. 1). A significant number of scholars have endeavored to characterize this phenomenon, delineate its dimensional parameters, and propose models and frameworks for its comprehension. According to Bennett and Gabriel (1999), the concept pertains to the capture, storage, dissemination, and utilization of knowledge. Additionally, Islam et al. (2011, p. 382) defined the concept as "process of creating, acquiring, capturing, manipulating, storing, disseminating and re-using knowledge both tangible and intangible knowledge assets available in implicit and explicit knowledge." According to Ogunbanwo et al. (2019), this approach is conducive to enhancing performance and fostering innovation in tertiary institutions. In essence, knowledge management can be defined as a process that assists enterprises in identifying, selecting, arranging, extending, and transferring important information and specialist knowledge (Hron, 2006).

As would be anticipated, knowledge management—defined as the acquisition, storage, and dissemination of knowledge—emerges as a pivotal managerial priority within any given organization (Beiryaei & Jamporazmay, 2010). This phenomenon has garnered considerable attention within the domains of health and medical science as well. This is predicated on the continuous growth of medical information, the fact that its utilization can significantly affect treatment and health outcomes, and the fact that it remains severely underutilized when needed (Abidi, 2007). Knowledge management in health can therefore benefit this area of science by helping to cope with the expansion of knowledge. The management of health-related knowledge is pivotal in determining how healthcare systems address the inundation of medical information, facilitating its effective reception, processing, and dissemination. This capacity enables healthcare systems to overcome the deluge of medical and health information (Candy, 2010) and to "promote and provide optimal, timely, effective, and pragmatic healthcare knowledge to healthcare professionals (and even to patients and individuals) where and when they need it to help them make high-quality, well-informed, and cost-effective patient care decisions" (Abidi, 2007, p. 2). The effective

management of health knowledge is of paramount importance, as it can ensure the quality of healthcare services provided by the health sector. Additionally, this approach would enhance individual decision-making processes, as the decision-making process itself is significantly influenced by knowledge (Morr & Subercaze, 2010).

In summary, a substantial body of literature exists, comprising numerous academic contributions that offer comprehensive overviews of various aspects of health knowledge management. Consequently, the substantial body of articulated accomplishments in this theme renders the analysis of research on the accomplished contributions beneficial in general. Such analysis enables researchers and academia to identify fundamental influences and obtain a well-structured overview of the characteristics and any developments in this research area. Bibliometric analysis is a research method employed to study trends in academic research. It involves the analysis of published scholarly works from databases such as Scopus or Web of Science. This approach offers insights into the global research landscape in a particular field, based on publication outputs (Alsharif et al., 2020). It has emerged as a significant methodology for analyzing dedicated research in a research field, providing academics with the following capabilities: (1) obtaining a comprehensive perspective, (2) identifying knowledge gaps; (3) generating novel research ideas, and (4) articulating their planned contributions to the area (Donthu et al., 2021). The objective of this study is to provide a comprehensive overview of the extant literature on health knowledge management. A bibliometric analysis of the field will be conducted to map out the existing perspectives and visualize the contributions made. This approach will facilitate a more profound comprehension of the state of the art and identify areas for future research.

## 2 METHODOLOGY

This study employs a cross-sectional descriptive method with a scientometric approach. Furthermore, data science principles play a crucial role in managing and analyzing the extensive bibliographic data collected from Web of Science and PubMed. The process entails a series of data science techniques, including data

cleaning, normalization, and structured processing. These techniques are imperative for guaranteeing the accuracy and consistency of the data. These techniques enable the efficient management of large datasets, facilitating the extraction of meaningful insights and patterns from thousands of research documents. This study employs a co-occurrence method to extract information from citation databases. Initially, it identifies relationships and patterns within the processed data. In the subsequent stage, the relationships are visualized using the VOSviewer tool. The co-occurrence method can be regarded as both a form of content analysis and, to a certain extent, a data mining technique. The utilization of data visualization tools, such as VOSviewer, serves to illustrate fundamental data science practices, including network analysis and graphical representation of intricate information. These tools facilitate the generation of co-occurrence maps and clusters, thereby elucidating thematic relationships and collaborative networks within the domain of health knowledge management. Integrating data science methodologies enhances the robustness and depth of the bibliometric analysis, ensuring a systematic approach to mapping research trends and providing actionable insights into the evolving landscape of health knowledge management.

### *2.1 Data collection*

The data for this study were retrieved from the Web of Science and PubMed databases on February 10, 2024. To ensure comprehensive coverage within the field, the study conducted searches using the following strategy: *((((Health knowledge management [Title/Abstract]) OR (Health km [Title/Abstract])) OR (*knowledge management [Title/Abstract])) OR (knowledge management [Title/Abstract])) AND (Health [Title/Abstract])*. Subsequently, a time span limitation was applied to the results, ranging from 1990 to 2023. This resulted in a total of 1,444 papers being reviewed. The data were stored in plain text format, and all files were aggregated into a single file.

### 2.2 Data analysis

The Web of Science analysis section was employed for the analysis, and Excel version 2019 software was subsequently utilized to create tables. The VOSviewer 1.6 software was utilized to generate science maps and co-occurrence maps, as well as to identify scientific clusters and newly formed co-occurrence clusters. In this visualization, the size of the circles is indicative of the weight, based on the co-occurrence of subjects, with a maximum length of 30. The data normalization method employed for network visualization was the minimum strength and association strength. During the course of the data review process, extraneous data were eliminated. It is imperative to acknowledge that the analyzed samples were scientific documents, including research articles, reviews, and books. Consequently, this study does not encompass ethical considerations.

## 3 RESULTS

A systematic analysis of the collected data was conducted using VOSviewer 1.6 and Excel software to identify key trends and insights in the domain of health knowledge management. The subsequent sections offer a thorough examination of the data, emphasizing notable findings and identifying recurring patterns. The period of greatest document production in the domain of health knowledge management was observed in 2022. The data indicate a consistent upward trend in scientific output from 1990, with a particularly significant surge between 2017 and 2022. This substantial increase underscores the growing recognition and importance of this field during these years. The health care sciences services subfield has demonstrated a leading role in scientific production, with a cumulative total of 277 records. It is noteworthy that the nascent field of medical informatics is positioned fourth, contributing 12% of the total scientific output. This underscores the mounting significance of knowledge management within the healthcare sector (Table 1).

**Table 1.** Research fields that have had the most scientific production in this field.

| Research areas | Record count |
|---|---|
| Health care sciences services | 277 |
| Business economics | 276 |
| Computer science | 264 |
| Medical informatics | 203 |
| Information science library science | 196 |
| Engineering | 160 |
| Public environmental occupational health | 164 |
| Operations research management science | 62 |
| General internal medicine | 61 |
| Environmental sciences ecology | 59 |

Wright A and Sittig DF have been identified as the most prolific authors in the field of health knowledge management from 1990 to 2024. The co-authorship network within this field has been revealed to consist of two distinct clusters. The first cluster consists of five authors centered around Wright D, while the second cluster comprises three authors centered around Middleton. This co-authorship map underscores the collaborative nature of research within this field. This study analyzes the scientific production in health knowledge management across various countries. The United States has accumulated an impressive number of records, with a total of 336, reflecting its leading role in the field. England follows with 158 records, and Canada is third with 112 records, both showing strong engagement. Other countries that have contributed significantly to this field include Australia, with 81 records, and Italy, with 89 records. Significant contributions have also been demonstrated by countries such as China (97 records) and Spain (91 records), suggesting a growing interest in health knowledge management. Among the countries exhibiting

lower outputs, Colombia has the fewest number of records at 20, while Iran, with 57 records, is notable for its contributions relative to its size and resources.

Figure 1 presents a scientometric map that illustrates international collaboration in the field of health knowledge management. The map is based on publication years and uses clustering to demonstrate the international collaborative efforts in this field. The map indicates that the United States and Canada have been consistently active contributors, demonstrating a long-standing tradition of producing scientific literature in this area. It is noteworthy that Iran and India have exhibited a marked increase in research output following 2019, underscoring their escalating engagement in health knowledge management research. This increase signifies a heightened global interest in the field, with these countries assuming a leading role in research initiatives.

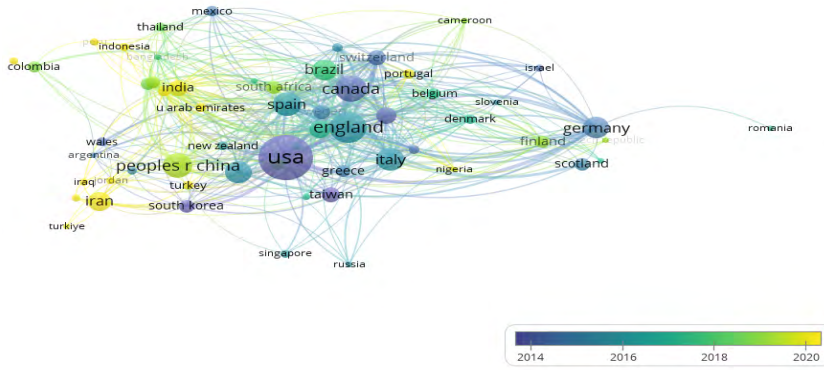


**Figure 2.** Scientometric map of cooperation between countries.

Figure 2 illustrates the prominent institutions contributing to research in health knowledge management. The scientometric map clearly demonstrates that Harvard University, the University of Toronto, and the World Health Organization (WHO) are at the forefront of significant research activities in this field. These institutions are recognized for their substantial contributions and leadership in advancing health knowledge management.

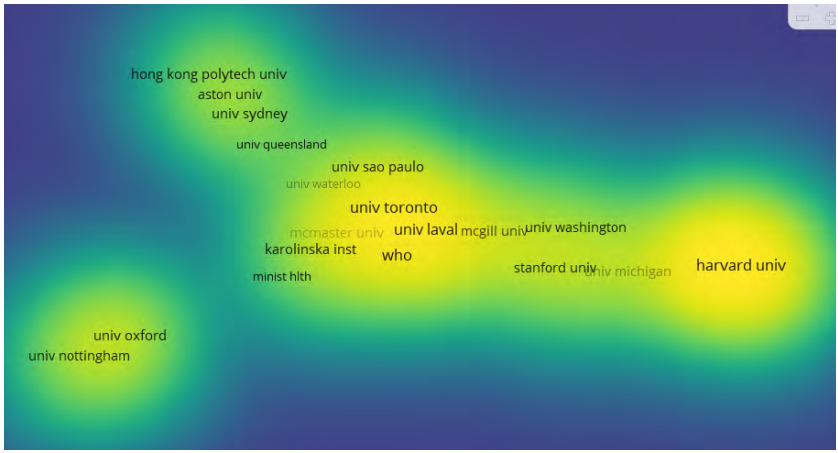


**Figure 3.** Institutions producing scientific records.

Subsequent findings demonstrate the distribution of research funding among prominent organizations in the domain of health knowledge management. The United States Department of Health and Human Services is at the forefront, with 56 documented allocations, signifying a pronounced dedication to promoting research in this domain. The Human Services sector has demonstrated notable support, with 45 recorded funds. It is noteworthy that the National Institutes of Health (NIH) in the United States plays a pivotal role in this regard, as evidenced by the 27 recorded funds, which underscore its crucial position in fostering innovation and research. The National Library of Medicine (NLM) of the National Institutes of Health (NIH) received 26 funds, underscoring its pivotal role in knowledge dissemination. On an international level, the National Natural Science Foundation of China (NSFC) has 23 recorded funds, thereby demonstrating China's active involvement in global health research. The Canadian Institutes of Health Research (CIHR) and the European Union (EU) have also made notable contributions, with 20 and 18 funds, respectively. The UK Research and Innovation (UKRI), the Spanish Government, the Agency for Healthcare Research and Quality (AHRQ), and the National Institutes of Health Research (NIHR) have provided substantial support, with each contributing between 10 and 13 funds (Table 2).

**Table 2.** Institutions that have provided the most research funds in this field.

| Institutions | Record count |
|---|---|
| United States Department of Health and Human Services | 56 |
| Human Services | 45 |
| National Institutes of Health (NIH) USA | 27 |
| NIH National Library of Medicine (NLM) | 26 |
| National Natural Science Foundation of China (NSFC) | 23 |
| Canadian Institutes of Health Research (CIHR) | 20 |
| European Union (EU) | 18 |
| UK Research and Innovation (UKRI) | 10–13 |
| Spanish Government | |
| Agency for Healthcare Research and Quality (AHRQ) | |
| National Institutes of Health Research (NIHR) | |

Figure 3 presents a scientific map illustrating the co-occurrences within the field, categorized by year. The map employs a color-coding system to differentiate between distinct time periods: purple signifies co-occurrences from 2015 and earlier, while yellow indicates the most recent occurrences. The map underscores several subjects that have gained prominence in recent years (from 2019 onwards), including big data, social media, innovation, and leadership, thereby emphasizing their growing significance in health knowledge management. Furthermore, since 2017, there has been an escalating focus on clinical decision-making and electronic health records, signifying an evolution in the emphasis on leveraging data and technology to optimize healthcare outcomes.

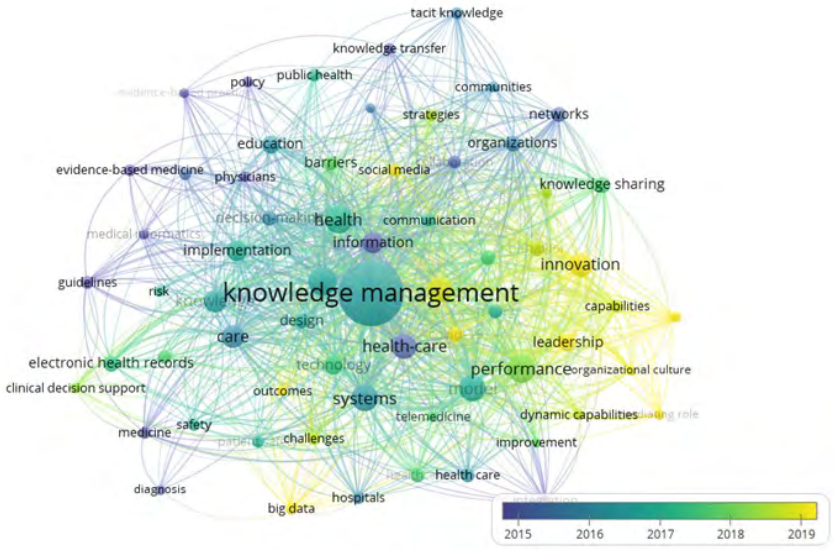


**Figure 4.** Scientometric map (general) of co-occurrences in the field of health knowledge management.

As illustrated in Figure 4, the general map of co-occurrences is presented, based on the most frequent co-occurrences, without consideration of the year. The map is divided into four distinct clusters, each represented by a different color. These colors highlight various words and phrases that are directly related to specific areas within the field of health knowledge management. The green cluster centers on the repercussions of the Coronavirus disease 2019 (COVID-19) on health knowledge management, accentuating the pivotal function of knowledge mobilization in addressing health crises. This cluster reflects how the pandemic has influenced the management and utilization of health information in emergency situations.

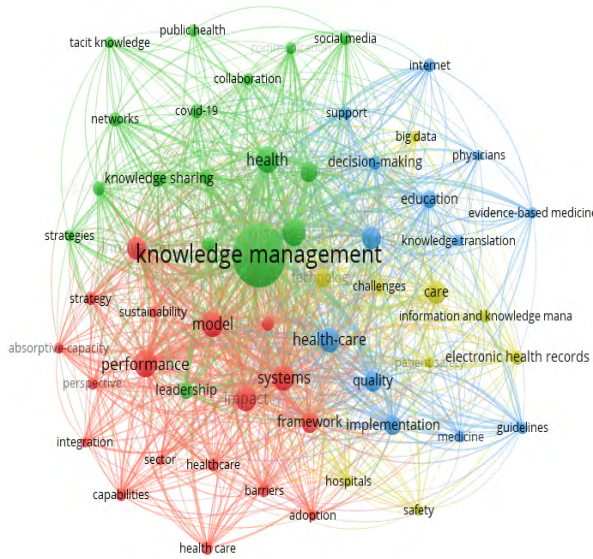


**Figure 5.** The general map of co-occurrences based on the most co-occurrences and regardless of the year of document production.

The red cluster centers on themes such as performance improvement, system activities, process optimization, resource management, and information systems. This underscores the necessity of enhancing efficiency and effectiveness within health systems through the implementation of superior management practices. This cluster underscores the ongoing endeavors to enhance health processes and resources, with the objective of optimizing overall system performance. The yellow cluster underscores the significance of electronic health records, hospitals, and the associated challenges. The text addresses subjects related to the enhancement of health services and the management of knowledge within electronic systems. This cluster underscores the pivotal function of digital health records in optimizing service delivery and patient outcomes, while concurrently addressing the obstacles encountered during the implementation and management of these systems. Consequently, the blue cluster places a pronounced emphasis on research and development, particularly

in domains such as big data, the Internet, fundamental research, and technological development. This development signifies the advancement of knowledge and technology in the health sector, illustrating the integration of innovative tools and data analytics in health research. The map offers a comprehensive representation of the multifaceted subjects related to knowledge management in health, illustrating the interconnection and contribution of diverse domains to the field's overarching understanding.
As illustrated in Figure 5, the initial cluster of the scientific map of co-occurrences in the domain of health knowledge management is represented. This cluster encompasses 19 pivotal items that play a substantial role in this domain. Among the most salient of these developments are the concepts of "big data," "electronic health records," and "patient safety," which have become pivotal components in the contemporary healthcare landscape.

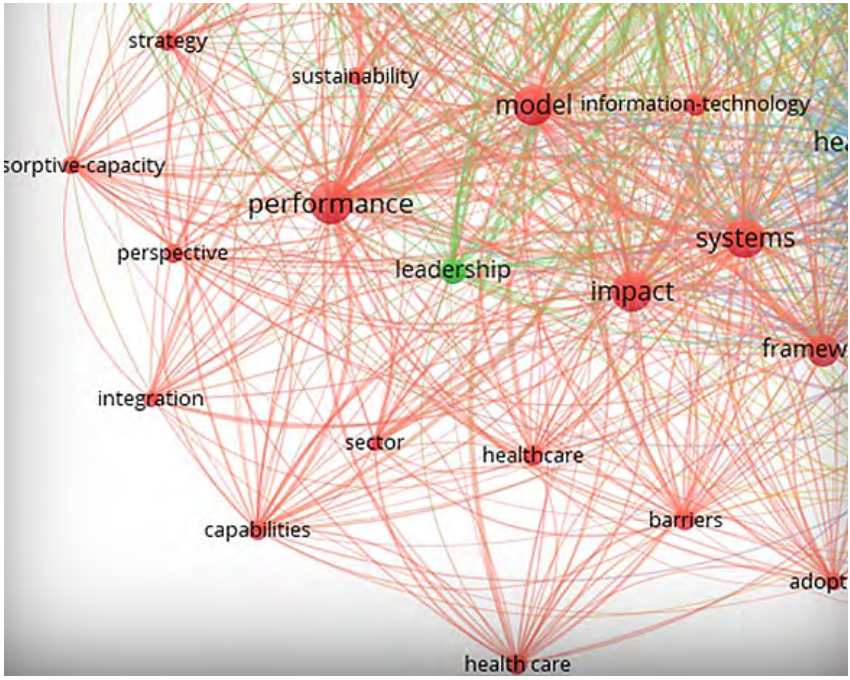


**Figure 6.** The first cluster of the scientometric map of co-occurrences in the field of health knowledge management.

As illustrated in Figure 6, the second cluster in the scientific map of co-occurrences in health knowledge management is presented. This cluster encompasses 19 pivotal elements, with a concentration on critical domains such as performance, dynamic capabilities, healthcare, information technology, and knowledge sharing.

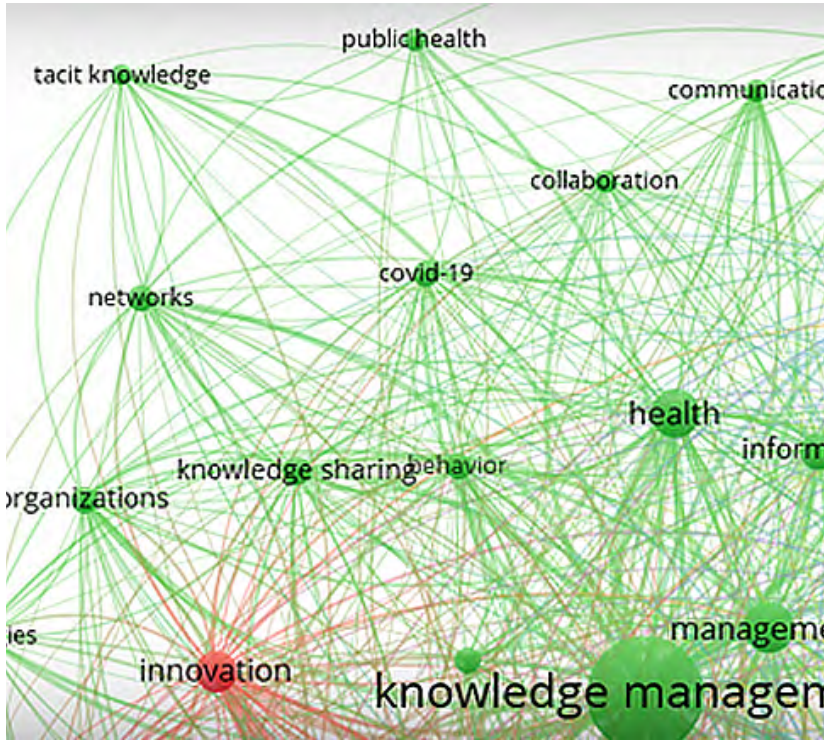


**Figure 7.** The second cluster of the scientometric map of co-occurrences in the field of health knowledge management.

As illustrated in Figure 7, the third cluster in the domain of health knowledge management comprises 13 items. This cluster places significant emphasis on decision-making, education, evidence-based medicine, and knowledge translation, underscoring the significance of informed decision-making and the dissemination of knowledge in healthcare.

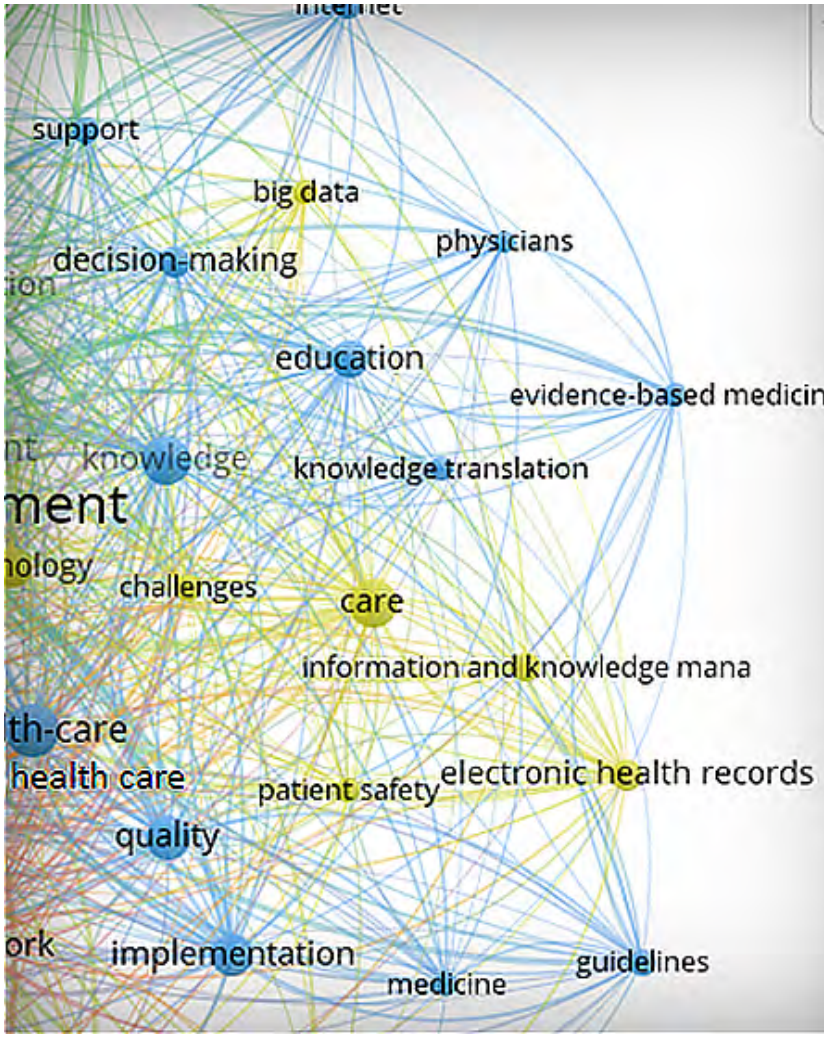


**Figure 8.** The third cluster of the scientometric map of co-occurrences in the field of health knowledge management.

Figure 8 illustrates the fourth and final co-occurrence cluster in the domain of health knowledge management. This cluster, which is comprised of nine items, focuses on topics such as big

data, electronic health records, and hospitals. It emphasizes the role of data and technology in modern healthcare systems.


**Figure 9.** The fourth cluster of the scientometric map of co-occurrences in the field of health knowledge management.

## 4 DISCUSSION

The bibliographic analysis conducted in this study illuminates various aspects of research in health knowledge management, offering valuable insights into the current state of the field and its implications for healthcare practice and policy. The analysis yielded several noteworthy trends in research output, including the preeminence of certain subfields within health knowledge management and the contributions of specific authors, institutions, and countries. Understanding these trends can inform future research priorities (Drysdale et al., 2013) and collaborations, helping to advance knowledge and innovation in healthcare. The present findings are consistent with those of previous research studies (Amri & Abed, 2023; Ibeh et al., 2024; Khang et al., 2024; Okolo et al., 2024; Ravikumar et al., 2023), which have also

identified emerging topics such as big data, social media, innovation, and leadership as key areas of focus within health knowledge management. These trends underscore the evolving nature of healthcare delivery and the increasing importance of leveraging technology and data-driven approaches to improve patient care and organizational efficiency. Bendowska and Baum's (2023) research demonstrated that medical care cooperation can result in several notable benefits, including enhanced patient safety, reduced hospitalization rates, and a decline in medical errors. Additionally, these findings suggest that such cooperation can also lead to improved patient access to medical services. The collaboration network that was revealed by the analysis highlights the importance of global cooperation in advancing research and addressing complex healthcare challenges. The United States, Canada, Iran, and India have emerged as key contributors to the field, indicating a growing interest and investment in health knowledge management across diverse geographic regions.

The findings of this study have practical implications for healthcare practice and policy. By identifying knowledge gaps and areas of research focus, healthcare organizations and policymakers can better allocate resources and prioritize initiatives aimed at enhancing knowledge management practices. This encompasses investments in information technologies, cultivation of a culture of knowledge sharing and collaboration, and formulation of strategies to effectively utilize health data for decision-making and quality improvement. While the analysis provides valuable insights into the current landscape of health knowledge management, several challenges and opportunities for future research warrant consideration. These objectives may encompass addressing issues related to data privacy and security, enhancing interdisciplinary collaboration between healthcare and information science disciplines, and exploring innovative approaches to knowledge dissemination and implementation in real-world healthcare settings. It is imperative to acknowledge the limitations of this study, including the reliance on bibliometric data, which may not fully capture the scope of research activities in health knowledge management. Furthermore, the analysis may be subject to biases inherent in the selection and interpretation of data, albeit unintentionally. Future research endeavors could augment bibliometric analysis with qualitative methodologies to facilitate a more comprehensive understanding

of the factors that influence research trends and outcomes in this domain.

## 5 CONCLUSION

In summary, this bibliographic analysis offers valuable insights into research trends in health knowledge management. The findings of this study underscore the significance of implementing effective knowledge management strategies, particularly in the context of technological advancements and global collaboration. Practical implications of this paradigm shift include the need for prioritizing resources, investing in information technologies, and fostering collaboration to improve healthcare outcomes. While this study offers valuable insights, future research should address its limitations and explore complementary methodologies. This analysis contributes to the understanding of health knowledge management and informs future research and practice in healthcare.

### *Funding*

This study received support from Abadan University of Medical Sciences (code: 1796).

### *Conflict of interest*

The authors declare that there are no conflicts of interest related to this research.

### *Contribution statement*

**Conceptualization:** Hossein Ghalavand
**Methodology:** Reza Varmazyar, Saied Shirshahi
**Data Curation:** Saied Shirshahi
**Formal Analysis:** Saied Shirshahi
**Writing – Original Draft:** Reza Varmazyar, Saied Shirshahi

**Writing – Review & Editing:** Hossein Ghalavand, Saied Shirshahi, Reza Varmazyar
**Supervision:** Hossein Ghalavand

### *Statement of data consent*

The bibliographic data used in this study, including data from Web of Science and PubMed, have been processed and analyzed as described in the methodology. The datasets generated during this research are available upon request and can be provided to interested researchers for further review.

## REFERENCES


Abidi, S. S. R. (2007). Healthcare knowledge management: The art of the possible. In *AIME workshop on knowledge management for health care procedures* (pp. 1–20). *https://doi.org/10.1007/978-3-540-78624-5_1*

Alsharif, A. H., Salleh, N., & Baharun, R. (2020). Bibliometric analysis. *Journal of Theoretical and Applied Information Technology, 98*(15), 2948–2962.

Amri, M. M., & Abed, S. A. (2023). The data-driven future of healthcare: A review. *Mesopotamian Journal of Big Data, 2023*, 68–74. *https://doi.org/10.58496/MJBD/2023/010*

Beiryaei, H. S., & Jamporazmay, M. (2010). Propose a framework for knowledge management strategic planning (KMSSP). In *2010 International conference on electronics and information engineering* (vol. 2, pp. V2-469–V2-473). *https://doi.org/10.1109/iceie.2010.5559819*

Bendowska, A., & Baum, E. (2023). The significance of cooperation in interdisciplinary health care teams as perceived by polish medical students. *International Journal of Environmental Research and Public Health, 20*(2), Article 2. *https://doi.org/10.3390/ijerph20020954*

Bennett, R., & Gabriel, H. (1999). Organisational factors and knowledge management within large marketing departments: An empirical study. *Journal of Knowledge Management, 3*(3), 212–225. *https://doi.org/10.1108/13673279910288707*

Candy, P. C. (2010). *Healthcare knowledge management: Issues, advances and successes*. Springer Science & Business Media.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research, 133*, 285–296. *https://doi.org/10.1016/j.jbusres.2021.04.070*

Drysdale, J. S., Graham, C. R., Spring, K. J., & Halverson, L. R. (2013). An analysis of research trends in dissertations and theses studying blended learning. *The Internet and Higher Education, 17*, 90–100. *https://doi.org/10.1016/j.iheduc.2012.11.003*

Hansen, M. T., Nohria, N., & Tierney, T. (2005). What's your strategy for managing knowledge. *Knowledge Management: Critical Perspectives on Business and Management, 77*(2), 322.

Hron, J. (2006). Knowledge and strategic management. *Agricultural Economics (Zemědělská Ekonomika), 52*(3), 101–106. *https://doi.org/10.17221/5001-AGRICECON*

Ibeh, C. V., Elufioye, O. A., Olorunsogo, T., Asuzu, O. F., Nduubuisi, N. L., Daraojimba, A. I., Ibeh, C. V., Elufioye, O. A., Olorunsogo, T., Asuzu, O. F., Nduubuisi, N. L., & Daraojimba, A. I. (2024). Data analytics in healthcare: A review of patient-centric approaches and healthcare delivery. *World Journal of Advanced Research and Reviews, 21*(2), Article 2. *https://doi.org/10.30574/wjarr.2024.21.2.0246*

Islam, M. S., Kunifuji, S., Miura, M., & Hayama, T. (2011). Adopting knowledge management in an e-learning system: Insights and views of KM and EL research scholars. *Knowledge Management & E-Learning, 3*(3), 375.

Khang, A., Triwiyanto, Abdullayev, V., Ali, R. N., Bali, S. Y., Mammadaga, G. M., & Hafiz, M. K. (2024). Using big data to solve problems in the field of medicine. In *Computer vision and AI-integrated IoT technologies in the medical ecosystem* (pp. 407–418). CRC Press.

Morr, C. E., & Subercaze, J. (2010). Knowledge management in healthcare. In *Handbook of research on developments in e-health and telemedicine: Technological and social perspectives* (pp. 490–510). IGI Global Scientific Publishing. *https://doi.org/10.4018/978-1-61520-670-4.ch023*

Ogunbanwo, A. S., Okesola, J. O., & Buckley, S. (2019). Knowledge management awareness assessment in Nigerian tertiary institutions. *F1000Research, 8*, 608. *https://doi.org/10.12688/f1000research.18223.2*

Okolo, C. A., Chidi, R., Babawarun, O., Arowoogun, J. O., Adeniyi, A. O., Okolo, C. A., Chidi, R., Babawarun, O., Arowoogun, J. O., & Adeniyi, A. O. (2024). Data-driven approaches to bridging the gap in health communication disparities: A systematic review. *World Journal of Advanced Research and Reviews*, *21*(2), Article 2. *https://doi.org/10.30574/wjarr.2024.21.2.0591*

Pandey, K. N. (2016). *Paradigms of knowledge management*. Springer.

Ravikumar, R., Kitana, A., Taamneh, A., Aburayya, A., Shwedeh, F., Salloum, S., & Shaalan, K. (2023). The impact of big data quality analytics on knowledge management in healthcare institutions: Lessons learned from big data's application within the healthcare sector. *South Eastern European Journal of Public Health*. *https://doi.org/10.56801/seejph.vi.309*


pro-metrics