CBIS'24
XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Uma análise bibliométrica de 50 anos de pesquisa em interoperabilidade

A 50-year bibliometric analysis of the interoperability research field

Un análisis bibliométrico de 50 años de investigación sobre interoperabilidad

Jackson Silva[1], Fagner Silva[2], Elyda Freitas[3], Everton R. Silva[4] , Cynthia Vieira[5], Vinicius Cardoso Garcia[6]

1 Prof. Msc., Campus Caruaru, Universidade de Pernambuco, Caruaru (PE), Brasil.
2 Doutorando, Centro de Informática, Universidade Federal de Pernambuco, Recife (PE), Brasil.
3 Profa. Dra., Campus Caruaru, Universidade de Pernambuco, Caruaru (PE), Brasil.
4 Graduando, Campus Caruaru, Universidade de Pernambuco, Caruaru (PE), Brasil.
5 Graduanda, Autarquia Educacional do Belo Jardim – Faculdade do Belo Jardim, Belo Jardim (PE)
6 Prof. Dr., Centro de Informática, Universidade Federal de Pernambuco, Recife (PE), Brasil.

Autor correspondente: Prof. Msc. Jackson Raniel F. da SIlva
E-mail: jackson.florencio@upe.br

Links: https://doi.org/10.6084/m9.figshare.26236574.v1

Resumo

Interoperabilidade, a operação coordenada de sistemas, é um conceito complexo com definições variadas. Objetivo: Este estudo investiga a evolução da pesquisa em interoperabilidade nos últimos 50 anos. Método: A avaliação de metadados biliométrica foi utilizada como método. A análise abrange 21.431 publicações. Resultados: Os resultados indicam uma contribuição significativa da ciência da computação, engenharia e matemática. Conclusão: Os resultados sugerem uma alta dispersão das pesquisas sobre interoperabilidade por fontes de publicações e áreas de conhecimento e uma predominância de propostas de soluções pontuais em diferentes contextos onde a interoperabilidade é apenas uma característica, em oposição às pesquisas conceituais sobre interoperabilidade.

Descritores: Bibliometria; Bases de Dados Bibliográficas; Interoperabilidade

Abstract

Interoperability, the coordinated operation of systems, is a complex concept with varying definitions. Objective: This study delves into the evolution of interoperability research over the past 50 years. Method: The method was the bibliometric metadata analysis. The analysis encompasses 21,431 publications. Results: The results indicate a significant contribution from computer science, engineering, and mathematics. Conclusions: The results suggest a high dispersion of research on interoperability by publications sources and areas of knowledge and a predominance of punctual solution proposals in different contexts where interoperability is just a feature, in opposition to conceptual research on interoperability.

Keywords: Bibliometrics; Bibliographic Databases; Interoperability

Resumen

Interoperabilidad, la operación coordinada de sistemas, es un concepto complejo con diferentes definiciones. Objetivo: Este estudio profundiza en la evolución de la investigación sobre interoperabilidad durante los últimos 50 años. Método: Se utilizó como método la evaluación de metadatos biliométricos. El análisis engloba 21.431 publicaciones. Resultados: Los resultados indican una contribución significativa de la informática, la ingeniería y las matemáticas. Conclusión: Los resultados sugieren una alta dispersión de las investigaciones sobre interoperabilidad por fuentes de publicaciones y áreas de conocimiento y un predominio de propuestas de soluciones puntuales en diferentes contextos donde la interoperabilidad es sólo una característica, en oposición a las investigaciones conceptuales sobre interoperabilidad.

Descriptores: Bibliometría; Bases de Datos Bibliográficas; Interoperabilidad

Introduction

Technological development has allowed the emergence of new applications and systems. These system parts integrate, interoperate, interconnect, intercommunicate, relate, or compose while aiming for a common objective. In an attempt to describe the relationship of a system's parts, some of these terms can be used as synonyms for communicative abilities and exchanges in certain situations, as subsets in others,[(1)] as an intersection of areas,[(2)] or as part of the solution to the same problem.[(3)]

These potentially conflicting definitions of interoperability within a specific area or across different knowledge areas make understanding these concepts harder, configuring the problem to be addressed by this paper. The meaning of the word interoperability is subject to be investigated to provide a common understanding that, as stated by Diallo:(4) does not fall into an infinite recursion; meets the necessary and sufficient requirements for interoperability; precisely defines what data, information, useful information, and context is; be able to explain interoperability as part of a system versus interoperability with respect to other systems.

As previously demonstrated,(5) interoperability has several definitions and classification types. This study considers mapping the knowledge areas interested in Interoperability a necessary step toward understanding that term. So, our objective is to find areas, documents, and people capable of contributing to the understanding of the interoperability meaning, by conducting bibliometric analysis.

The remainder of this paper is organized as follows: the section Method describes the process and tools employed in this research; then, the section Results and Discussion brings the answers to the research questions raised. Finally, the conclusions present the knowledge obtained by carrying out this research, as well as its limitations.

Method

The methodological outline of this study follows the Zupic and Cater(6) workflow recommendations, containing the following steps: research planning, bibliometric data compiling, analysis, visualization, and interpretation. And, the approach to analyzing a research field proposed by Cobo(7) was also employed. Furthermore, we used the Marcos-Pablos and García-Peñalvo(8) string-builder tools and methods. Bibliometrics methods are suitable for quantitatively reviewing the literature. These methods enable literature reviews with thousands of documents by filtering their metadata.

The research questions for this study are: Which papers influenced the interoperability research the most? And, which publication sources and knowledge areas had the most significant influence on interoperability research? The next research steps description are in the following sections.

Data Compiling

The search string definition process minimizes the bias of the words known by the researcher and searches for new keywords to compose the scope of the research.[(8)] The planning of the relevant research terms contains five steps: 1. A pilot search based on the researcher's prior knowledge; 2. Exclusion of duplicate papers and papers without abstracts; 3. Search results classification according to their relevancy; 4. Use the statistical method Term Frequency - Inverse Document Frequency (TF-IDF) in the group of relevant articles to prospect new keywords; 5. Repeat the entire process until no new keywords appear.

Marcos-Pablos and García-Peñalvo[(8)] developed a Python script using the natural language processing libraries nltk, scitik-learn, and pandas that support this classification and new keywords suggestion process. The classification occurred with multinomial classifier Naive Bayes - Bigram frequencies with cross-validation of 10 folds. The F1-score calculated as a form of quality verification of the classification was 0.784. This result was the best among the classifiers Bernoulli Naive Bayes, Multinomial Naive Bayes, KVN, and SVN.

It was also necessary to remove from the results documents of the editorial type (ed), review (re), conference review (cr), letter (le), erratum (er), and notes (no), as well as withdrawing documents that are not direct intervention reports containing focus groups and bibliometric and scientometric analyses. The resulting search string was:

> TITLE-ABS-KEY ((system* OR data* OR model* OR information OR process OR cloud OR platform OR architecture) AND (improv* OR development) AND (interoper*)) AND NOT ("focus group" OR bibliometric OR scientometric) AND (LIMIT-TO(PUB-YEAR, 2020)) AND (EXCLUDE(DOCTYPE, "re") OR EXCLUDE(DOCTYPE, "cr") OR EXCLUDE(DOCTYPE, "ed") OR EXCLUDE (DOCTYPE, "no") OR EXCLUDE(DOCTYPE, "le") OR EXCLUDE (DOCTYPE, "er"))

In this research, the raw data come from Web of Science (WoS) and Scopus databases. After the search, the classifier considered 31,225 articles as relevant. The abstract cosine similarity calculus was a way of checking the classification. It achieved the similarity rate by comparing each abstract with the first relevant abstract. Graphically, it is possible to verify the high degree of similarity between the abstracts in the distribution chart of Fig. 1 (A).

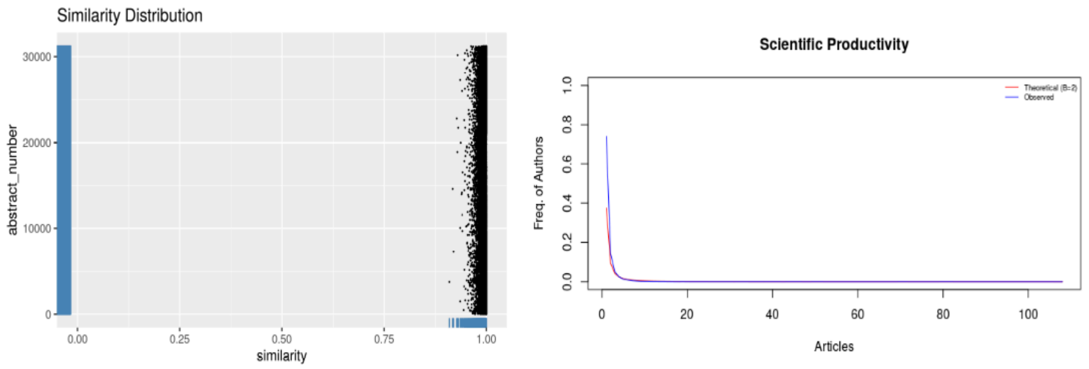
Considering the multiple sources of document metadata, it was necessary to agglomerate sets of subjects related to each other in 26 knowledge areas, maintaining the semantic content between the different databases. This classification in the study areas proposed was revised based on the Classification of Instructional Programs - 2020 (CIP) and is available as a supplementary resource by accessing the link provided in this paper.[9]

Pre-Processing

Data from bibliographic sources commonly contain errors[7], needing pre-processing. The downloaded data were divided by area of knowledge and the areas with many results represented more from a source file as a result of search engine. The first step was to import this data separated by areas and search engines for the construction of the data frames in R (version 3.6.3) utilizing the library Bibliometrix (Version 3.0.1).[10]

Subsequently, we merged data frames from the same knowledge area, even from distinct databases. In this process, the Bibliometrix[10] eliminated eventual duplicate source documents between search engines. This process resulted in 26 thematic data frames. We performed a second verification of duplicates utilizing the algorithm of distance by Damerau-Levenshtein [11] analyzing document titles with at least 95% similarity. This verification did not eliminate any document. We performed a third verification in search of misspellings, considering the title and publishing year in each thematic data frame.

Figure 1 - (A) Abstracts Similarity Distribution. Each point in the chart represents an abstract, and the distance between them is allusive to the calculated cosine distance. (B) Lotka's Theoretical distribution comparison.



(A) (B)

Results and discussion

The group of all documents comprehends publications made between 1970 and 2021, totaling 21,431 documents from 9,685 different publishing sources. Out of these, 7,095 are articles, 8,433 are conference papers, and 4,649 are proceedings papers. These documents represent the work of 47,048 authors, with only 3,010 documents single-authored. About 73.98% of the authors published only one document. This distribution can be perceived in the distribution chart of Lotka [12] in Fig. 1 (B).

It assumes an inverse square law in which the number of authors making a certain number of contributions is a fixed ratio to the number of authors publishing a single article, implying that the theoretical beta coefficient of Lotka's law nearly always equals 2. In the group of observed documents, we obtained a B = 2.37, a constant of 0.50 and a goodness of fit R2 = 0.94. However, the Kolmogorov-Smirnov test between the theoretical distributions and observed returned a p-value of $1.456891e^{-06}$, which rejects the null hypothesis.

It results in a bibliometric database with very similar papers when comparing their abstracts, as can be seen in the table of dispersion of Fig. 1. After being sanitized and compared with Lotka[12] theoretical distribution, this database showed great goodness of fit, which also shows up as an indication of the quality of the analyzed database.

Given that evidence, we considered that the usage of a semi-automated and systematic method of search string building reduced eventual bias from the experiences and expectations of the authors from this study. We also considered that this methodological approach is the first contribution of this research since it can be helpful to enable massive bibliometric analysis, reducing bias in the construction of the search strings in future research. Additionally, these results indicate that a significant group of authors published a solution to the problem relative to interoperability in their specific areas of knowledge and did not feel stimulated to contribute more deeply to the development of the interoperability research itself.

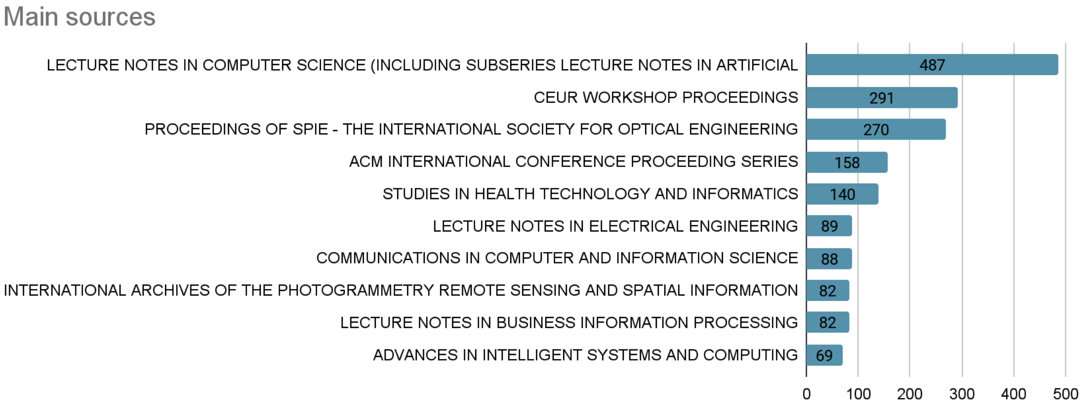
Which publication sources and knowledge areas had the most significant influence on interoperability research?

The research about interoperability is spread through 26 knowledge areas, as segmented in the supplementary resource by accessing the link provided in this paper. The on-line supplementary material presents the areas with the highest number of articles from 1970 to 2020. The volume of papers published in Computer Science and Engineering stands at 13,423 and 10,498, respectively. It represents about three times more than Mathematics, which appears in third place.

The majority of papers were published in the areas of Computer Science and Engineering, followed by Biological Sciences. Analogously, the second group of bars shows the number of interdisciplinary papers exclusively in two areas of knowledge, predominating the intersections between Computer Science and Engineering, Computer Science and Mathematics, and Computer and Social Sciences. Observing the third and fourth group of bars, the prevailing compound intersections are again the areas of Computer Science and Engineering.

Given the volume of publications per area, the expectation is that the sources of publications reflect the highlighted areas. This expectation is confirmed by Figure 2. However, the volume of publications from these sources represents only about 8% of the published papers. It is an indication that there are no journals or conferences that aggregate a significant number of publications in the area of interoperability. So, these observations allow us to conclude that the studies produced about interoperability potentially apply techniques of this area to problems of the other areas of knowledge.

Figure 2 - Main Sources



Which papers influenced the interoperability research the most?

To answer this research question, the top ten articles with the highest citation count were considered the most influential, classified in descending order, among the contemplated decades and by their publication areas. This way, considering the study's relevance in the decade that it was published and reducing the time bias affects the simple counting of citations, resulting in four 10-paper lists.

From these lists' appreciation, we intend to verify how the most influential papers have contributed to developing knowledge about interoperability and not classify these papers according to the type of interoperability they present, as done by Panetto(13) and Ford *et. al.*,(5) nor classify the contribution given by some level of interoperability. While these are good ways to compare contributions, these classifications serve specific purposes in specific domains. However, our data come from different domains, and we are more interested in the final products of the papers than in how they rank according to any other parameter.

Using the abstracts of the most cited articles, it was possible to classify the type of influence that they exert from their main contributions. Going from the analysis of the most cited articles by areas of knowledge, it is possible to notice a predominance of papers that present a solution that needs to have interoperability as a characteristic. Followed by papers that discuss interoperability from the point of view of a communication standard or data storage, and only three discuss the concept of interoperability directly. The other four papers explain the insertion of one part in a system (integration), and three papers discuss interoperability as something present in the system where a solution is executed.

In all knowledge areas, there is a predominance of papers that present a solution with interoperability as a characteristic, except for the areas of Business and Social Sciences that brought more conceptual discussions and papers about communication standards and data storage. The solutions with interoperability presented by the articles in the areas of Physical and Natural Sciences are, in their majority, tools that solve specific solutions from these areas. On the other hand, the interoperability solutions classified in Mathematics, Engineering, and Computer Science appear repeatedly, representing the construction of generic frameworks or software with applications in other areas of knowledge. In the case of classified papers in the Mathematics knowledge area, none of them address any mathematical formalism relative to interoperability, which leads us to believe that the analysis of returned articles from more than one area concurrently.

The analysis of the most cited articles throughout the time follows a similar tendency. Considering the most cited articles, there is a paper that treats interoperability as something present in the system in which a solution is executed and a paper that talks about a communication standard or data storage. The rest discuss solutions that have or need interoperability as a characteristic.

The six most cited papers approach conceptual discussions and have a publishing date before 2005. Even so, only the paper from Mensh, Kite, and Darby(14) approaches interoperability concepts directly. The other ones primordially approach other concepts that come into a relationship with interoperability. Of the other six papers, five of them were published between 1981 and 1990 and the most recent was published in 2007, approach communication standards and data storage, leaving interoperability in the background.

It is highlighted in the analysis of four papers that treat interoperability as something present where one solution runs. Two treat interoperability as a period of elapsed time between surgery interventions. The others show solutions that have or need interoperability as a characteristic.

These results endorse the suspicion from which many authors published a solution to a problem related to interoperability in their specific areas of knowledge and did not feel stimulated to contribute more profoundly to the development of the interoperability area itself, creating even multiple proposals of solutions inner the same application domains.[15] Moreover, the low number of papers that influenced the research about interoperability in a way to add theoretical or conceptual discussions indicates that the solutions created are not conceptually grounded, at least in a straightforward way, or that ground themselves in old and eventually out-of-date concepts.

Conclusions

This article presented a new methodological way for the composition of databases of bibliometric studies, presented an execution of an analysis of citations, and discussed characteristics of the interoperability area based on the results found. The citation analysis identified the publication sources and research areas that influenced the studies about interoperability, with the number of citations as a metric. In an analog way, the most influential papers were identified through an analysis of knowledge areas, with a timely analysis segmented by groups of years covering the period from 1970 to 2020.

The methodological framework utilized led us to interpret that the database analyzed: 1) Indicates that the research about interoperability is, in fact, interdisciplinary; 2) There is no predominance of publications in vehicles from a specific knowledge area; 3) Interoperability research studies are applications of interoperability techniques in problems of multiple areas of knowledge; 4) The analyzed authors tend to publish a solution to a problem related to interoperability with application in some specific knowledge area and do not feel stimulated to contribute in a more in-depth way to the interoperability itself; 5) There is a predominance of publications that treats interoperability as an existing feature or a desirable one in a system; 6) Business and Social Sciences are the knowledge areas where the conceptual discussions about interoperability predominate; 7) The concepts of interoperability found are from 15 to 41 years ago; 8) The proposition of solutions with interoperability is not grounded, at least explicitly, in concepts of interoperability or grounds themselves in old and potentially outdated concepts.

There is room for deeper analysis of the collected data, focusing on specific research areas or application domains or executing other bibliometric analyses. There are also research opportunities for developing automation tools for the search string-building process and tools that help sanitize the bibliometric database.

It is worth noting that the citation analysis focuses only on the most cited articles, and the sum of the articles less cited may be more significant than the influence of them.[16] There is the risk of bias by auto citations and citations "in-house"[17]. In some citations, the intention is to refute or critique the cited paper[6]. However, it is impossible to identify the motivations behind a citation[18]. These situations must be considered inherent limitations to the bibliometric studies.

We used the CIP codes for a different function from the original one. However, a classification parameter in knowledge areas was necessary, and the CPI classification fulfilled the role of classification for such finality in the absence of a more adequate one.

References

1. Thomas I, Nejmeh B. Definitions of tool integration for environments. IEEE Software. 1992;9(2):29-35. doi: 10.1109/52.120599.

2. Gürdür D, Asplund F. A systematic review to merge discourses: Interoperability, integration and cyber-physical systems. J Ind Inf Integr. 2018;9:14-23. doi: 10.1016/j.jii.2017.12.001.

3. Naudet Y, Latour T, Guedria W, Chen D. Towards a systemic formalisation of interoperability. Comput Ind. 2010;61(2):176-85.

4. Diallo SY, Herencia-Zapana H, Padilla JJ, Tolk A. Understanding interoperability. In: Proceedings of the 2011 Emerging M&S Applications in Industry and Academia Symposium. 2011. p. 84-91.

5. Ford T, Colombi J, Graham S, Jacques D. Survey on interoperability measurement. In: Proceedings of the 2007 Interoperability for Enterprise Software and Applications Conference. 2007 Jun; Gold Coast, Australia. p. 67.

6. Zupic I, Čater T. Bibliometric methods in management and organization. Organ Res Methods. 2015;18(3):429-72. doi: 10.1177/1094428114562629.

7. Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. J Informetr. 2011;5(1):146-66.

8. Marcos-Pablos S, García-Peñalvo FJ. Decision support tools for SLR search string construction. In: Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'18). New York, NY, USA: ACM; 2018. p. 660-7.

9. Department of Education U.S. The classification of instructional programs. 2020.

10. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. J Informetr. 2017;11(4):959-75.

11. Bard GV. Spelling-error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string-edit distance metric. In: Proceedings of the Fifth Australasian Symposium on ACSW Frontiers (ACSW '07). AUS: Australian Computer Society, Inc.; 2007. p. 117-24.

12. Lotka AJ. The frequency distribution of scientific productivity. J Wash Acad Sci. 1926;16(12):317-23.

13. Panetto H. Towards a classification framework for interoperability of enterprise applications. Int J Comput Integr Manuf. 2007;20(8):727-40. doi: 10.1080/09511920600996419.

14. Mensh DR, Kite RS, Darby PH. The Quantification of Interoperability. Nav Eng J. 1989;101(3):251-9.

15. Ferreira DE, de Souza JM. Methodology for developing openehr archetypes: a narrative literature review. J Health Inform. 2023;15(2):53-9.

16. Martin BR. The evolution of science policy and innovation studies. Res Policy. 2012;41(7):1219-39.

17. van Raan AFJ. Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. Scientometrics. 1996;36(3):397-420.

18. Coombes PH, Nicholson JD. Business models and their relationship with marketing: A systematic literature review. Ind Mark Manag. 2013;42(5):656-64.