Controlled vocabularies in scientific literature's indexing: The case of the 1918 pandemic

Montserrat Garcia Alsina
Universitat Oberta de Catalunya, Spain.
Email: mgarciaals@uoc.edu
ORCID: https://orcid.org/0000-0002-1825-2279

Josep Cobarsi-Morales
Universitat Oberta de Catalunya, Spain.
Email: jcobarsi@gmail.com
ORCID: https://orcid.org/0000-0002-4382-1058

ABSTRACT

The scientific interest in the 1918 flu pandemic has been reinforced by the emergence in the early 21st century of epidemic pneumonia diseases caused by a virus, and more recently, the emergence of the SARS-CoV-2 virus, which caused the global pandemic known as "COVID-19," in 2020. This paper presents the findings of an exploratory study on the use of controlled languages in the scientific community, with the aim of identifying the knowledge generated and needed. This research has two objectives. The first is to identify the relevant controlled languages used by the scientific community to label the knowledge produced. The second is to ascertain the role played by controlled vocabularies in the recovery of scientific production. The research is centered on the production of literature concerning the 1918 pandemic, which has been indexed in two widely utilized databases: Web of Science and Scopus. Additionally, the investigation encompasses the controlled vocabularies pertinent to medical and health sciences subjects. Following the identification of articles pertaining to the subject matter, the scientific journals from which the articles have been retrieved are selected. Subsequently, the paper examines the instructions and guidance provided to authors by the journals in question, with the objective of analyzing the role played by keywords and controlled vocabularies in the scientific literature with regard to indexing and recovering knowledge in scientific databases. The preliminary results indicate that controlled vocabularies are infrequently utilized by journal publishers, as they are not included in the instructions provided to authors.

Keywords: controlled vocabularies in health field, scientific information retrieval, 1918 pandemic, scientific edition, scientific authorship

How to cite: Garcia Alsina, M., & Cobarsi-Morales, J. (2024). Controlled vocabularies in scientific literature's indexing: The case of the 1918 pandemic. In A. Angeluci, J. C. Morales, S. M. Cardama, & D. L. Arias (Eds.), Spanish and Portuguese contributions to the iConference 2024, Hybrid event, Changchun, China, 15-18/22-26 April 2024, Proceedings. *Advanced Notes in Information Science*, *volume 7* (pp. 109-123). Tallinn, Estonia: Pro-Metrics. DOI: 10.47909/978-9916-9974-8-2.87

1. INTRODUCTION

Keywords are an essential tool for the representation of knowledge. They are employed in the context of scientific production for the purpose of storage and subsequent retrieval in scientific databases. The words are collated and stored in databases as metadata in the keyword field. Additionally, keywords are utilized to represent the content within the title, abstract, and body of the text, with this content subsequently undergoing automatic indexing. The insufficient utilization of keywords in the indexing of scientific literature pertaining to the 1918 pandemic has been observed (Barry, 2004; Garcia-Alsina & Cobarsí, 2022; Knobler et al., 2005; World Health Organization, 2015). We will discuss these points in more detail below. In light of the aforementioned context, the aim of this paper is to examine the languages that the scientific community deems most appropriate for indexing the knowledge produced concerning the 1918 pandemic, which is often erroneously referred to in colloquial language and in the scientific literature as the "Spanish flu."

The study of indexing, encompassing both automatic and manual approaches, along with the role of controlled languages and natural language, has been a prominent area of research (Anderson & Perez, 2001; Baeza-Yates & Ribeiro-Neto, 2011; Ghanbarpour & Naderi, 20; Harter, 1975a, 1975b; Hong et al., 2009; Ishida et al., 2020; Jahoda, 1970; Lancaster, 1968; Veyette, 1961). The use of author-provided keywords and controlled vocabularies is a topic of ongoing debate and study (White, 2013). The identification of appropriate keywords is a matter of contention, with some scholars advocating for the unrestricted selection of words by the author, while others favor automated extraction (Ghanbarpour & Naderi, 2019; Harter, 1975a, 1975b; Ishida et al., 2020; Kwon, 2018; Lu et al., 2020; Zhang, 2008). In particular, controlled languages (such as thesauri, ontologies, taxonomies, or lists of headings) facilitate the representation of knowledge by ensuring the univocity of meanings, taking into account existing polysemies and synonymies (Keyser, 2012; Leise, 2008).

The decision to utilize controlled languages, both in manual and automatic indexing, is initially left to the discretion of the publishers who disseminate the scientific literature and the database managers who oversee its storage. Secondly, in the event that authors are permitted to select their own keywords, they are confronted with a multitude of potential options. One such option is to select keywords without fully understanding their relevance for the article to be found and without employing a strategy to do so (Lu et al., 2020). This ultimately results in suboptimal indexing. Another avenue available to authors is the voluntary choice of controlled languages to identify the most pertinent words (Ishida et al., 2020).

In any case, the cost of automatic versus manual indexing leads publishers to prefer automatic indexing (Zhang, 2008), which may result in the selection of keywords being regarded as a secondary consideration. Similarly, studies have indicated that keywords created by authors are less efficient than those extracted automatically (White, 2013; White et al., 2012). Other studies indicate that both methods of indexing (human and automatic) can be combined, thereby extracting advantages from both (Anderson & Perez, 2001). Furthermore, keywords and controlled languages are currently gaining even more strength for automated indexing, especially for retrieval. However, automation still requires further development and the incorporation of controlled vocabularies (Ahmad et al., 2020; Golub, 2021).

Another factor that has been considered in the study of indexing is the disparate utilization of keywords across disciplinary boundaries. In this regard, prior research suggests a tendency for authors in different disciplines to utilize keywords in a less interdisciplinary manner (Kwon, 2018). In the context of literature pertaining to the 1918 pandemic, the term "Spanish flu" is not merely a colloquialism but is also employed in scientific discourse (Garcia-Alsina and Cobarsí, 2022). The utilization of geographical terms associated with diseases contravenes the recommendations set forth by the World Health Organization (2015). Furthermore, the findings of several studies have challenged the hypothesis that the 1918 pandemic originated in Spain (Barry, 2004). It is essential that controlled vocabularies achieve consistency between the description of the content and its subsequent retrieval, through proper integration into the database. This indicates a failure in the indexing of the scientific literature related to this topic in

the databases, particularly if we consider the use of different terms (including "Spanish flu") to retrieve a single concept such as "pandemic of 1918." It is also necessary to consider the treatment of this term in controlled languages, including generalist (Library of Congress Subject Headings or UNESCO) and specialized languages in the fields of health (Medical Subject Headings [MeSH]) and humanities and social sciences (HASSET). An exploration of the Basic Register of Thesauri, Ontologies and Classification (BARTOC) indicates the existence of specific terms linked to pandemic or influenza, which do not include the term "Spanish flu."

In essence, this study examines the instructions that scholarly journals provide to authors regarding the use of keywords and the framework they must adhere to for their work to be indexed. This research phase begins with the following question: What are the criteria that scientific journals recommend to authors for selecting the languages in which they label the knowledge they produce?

2. METHODOLOGY

The research is based on articles produced between 2000 and 2019 on the 1918 pandemic, which have been indexed in two databases. The databases utilized for this research are Web of Science (WoS) and Scopus. The choice of years is motivated by the interest that arose from 2003 onwards in this topic following the onset of the SARS epidemic that led to an increase in research prior to the COVID-19 pandemic. To identify relevant journals for fieldwork, a search was conducted using four keywords in English: "Spanish influenza," "Spanish flu," "1918 influenza," and "1918 flu." The selected terms included synonyms pertaining to the disease itself and the various forms in which it was

referred to, including country and year of occurrence. A total of 70 articles published in 61 journals were identified through these searches. The aforementioned journals were then located on their respective websites, where the publication guidelines for authors could be accessed. After examining the websites, we excluded certain journals from the study based on the following criteria: those that have published informative or discussion articles; those published in a language unknown to the authors of this study (Korean, Icelandic, Norwegian, and Swedish), as we were unable to identify the instructions for authors; and those that are no longer published, thus lacking access to the instructions that the authors had at that time. In total, our study was based on a list of 49 journals.

The following information was extracted from the instructions to authors:

a. The fields to which the journal belongs: health sciences, experimental sciences, computer engineering, social sciences, humanities, and interdisciplinary.
b. The existence of instructions to authors on how to select keywords.
c. The specification of whether a controlled vocabulary should be used or whether the terms to be used are of free creation.
d. The vocabulary to be used by the author, if applicable.
e. The indication linked to search engine optimization (SEO), if applicable.

3. RESULTS

A content analysis of the instructions for authors on the websites of academic journals reveals a preponderance of

journals in the health sciences, followed by those in the social sciences and humanities (Table 1).

Table 1. Thematic scope of the journals.

FIELD	NUMBER OF JOURNALS	PERCENTAGE (%)
Health sciences	29	59.18
Humanities	8	16.32
Social sciences	6	12.24
Experimental sciences	4	8.16
Computer engineering	1	2.04
Interdisciplinary	1	2.04

Source: Own elaboration.

A total of 59.18% of the journals examined provide instructions to authors. The majority of these journals (68.97%) are in the field of health sciences, while 10.34% are in the fields of humanities and social sciences and offer instructions to authors on keywords (Table 2).

Table 2. Scope of journals with instructions to authors.

FIELD	NUMBER OF JOURNALS	PERCENTAGE OF JOURNALS WITH INSTRUCTIONS (OUT OF THE TOTAL OF JOURNALS WITH INSTRUCTIONS) (%)
Health sciences	20	68.97
Social sciences	3	10.34

(*Continued*)

Table 2. *Continued*

FIELD	NUMBER OF JOURNALS	PERCENTAGE OF JOURNALS WITH INSTRUCTIONS (OUT OF THE TOTAL OF JOURNALS WITH INSTRUCTIONS) (%)
Humanities	3	10.34
Experimental sciences	2	6.90
Computer engineering	1	3.45
Interdisciplinary	0	0

Source: Own elaboration.

A significant proportion of journals (40.81%) still fail to provide any indication of keywords in their guidelines for authors seeking to publish, which suggests a lack of evaluation of keywords by these journals (Table 3).

Table 3. Existence of instructions on keywords.

INSTRUCTIONS ON KEYWORDS	NUMBER OF JOURNALS	PERCENTAGE OF JOURNALS (%)
Journals without instructions	20	40.81
Journals with instructions	29	59.18

Source: Author.

In the case of journals that provide guidance, the majority of keywords are left to the dis cretion of the author, with only a minority of journals suggesting the use of a controlled vocabulary. Consequently, the selection of

keywords and their corresponding indexing is at the discretion of the authors, which may result in content being difficult to retrieve in searches. Table 4 provides a summary of this aspect.

Table 4. Specification of instructions on keywords.

INSTRUCTIONS ON KEYWORDS	NUMBER OF JOURNALS	PERCENTAGE OF JOURNALS (%)
Keywords by free choice	20	69
Keywords by controlled vocabulary	9	31

Source: Own elaboration.

In the case of keywords freely chosen by the author, a common guideline refers to establishing a minimum and/or maximum number of keywords (23 of the 29 journals that offer instructions to authors do so). It is uncommon for other guidelines to be provided, except in some cases where advice is given regarding the use of terms that will facilitate the dissemination of articles. In regard to the journals that indicate the specific use of a controlled vocabulary, three languages stand out, two of which are in the field of health and one in the social sciences. In the field of health, the two most commonly used languages are MeSH and Cumulative Index to Nursing and Allied Health Literature (CINAHL). In the social sciences, the Journal of Economic Literature (JEL) classification system is the most prevalent. This system is utilized to classify scientific literature in the field of economics, as indicated in the guide for authors (Table 5).

Table 5. Use of controlled languages in journals with instructions.

VOCABULARY	FIELD	NUMBER OF JOURNALS	PERCENTAGE OF JOURNALS (IN RELATION TO THE ONES PROVIDING WITH INSTRUCTIONS) (%)	PERCENTAGE OF JOURNALS (IN RELATION TO THE TOTAL EXAMINED) (%)
MeSH	Health Sciences	6	66.66	12.24
MeSH and CINAHL	Health Sciences	2	22.22	4.08
JEL	Social Sciences	1	11.11	2.04

Source: Own elaboration.

Furthermore, when the total number of journals studied (49) is considered alongside the number of journals that utilize controlled languages (nine), it becomes evident that there is a notable lack of promotion of the representation and indexing of knowledge. Indeed, only 18.36% of the journals in question advocate for the use of controlled languages. Conversely, in a subset of journals (12.24%), there is a necessity to inform authors of the importance of keywords, not only in the dedicated keyword section but also in the title, abstract, and the body of the article itself. The recommendations are designed to enhance SEO, thereby facilitating the discovery of articles on the Internet, whether in Google Scholar or other open repositories. It is notable that none of the instructions pertain to the automatic indexing of publishers' databases. It is also noteworthy that the focus is on SEO rather than on the efficiency and quality of information retrieval, with the aim of eliminating noise and documentary silence. An analysis of the instructions reveals that the journals that value and emphasize SEO do so from the perspective of disseminating the authors' production and, therefore, that of the journal itself. The retrieval of information in a more relevant and comprehensive manner is not secondary to the importance of keywords given in these journals and their instructions.

A review of the available evidence suggests that keywords and, in particular, their optimization for indexing and retrieval are not a primary concern for authors submitting articles to journals. This is particularly noteworthy when compared to the more frequent and explicit requirements set forth in journal instructions to authors, such as formatting of bibliographic references and anti-plagiarism guidelines. Thus far, our analysis of journal instructions

has been limited to those published on open access websites. Consequently, we have not considered the forms and applications employed by many journals to collect submissions, which may contain supplementary instructions embedded in their interface.

4. DISCUSSION

The findings of this preliminary investigation indicate that the majority of scientific journal publishers do not utilize controlled languages to effectively represent, index, and retrieve knowledge. This is due to the fact that they do not include such languages in their instructions to authors. Consequently, they are unaware of the potential of these languages to eliminate the effects of irrelevant or missing information. This use appears to be inconsistent with the requirements of information retrieval systems and indexing tools, both automatic and manual, as identified by researchers in this field, as pointed out by some of the countless studies in the field (Anderson & Perez, 2001; Baeza-Yates & Ribeiro-Neto, 2011; Ghanbarpour & Naderi, 2019; Harter, 1975a, 1975b; Hong et al., 2009; Ishida et al., 2020; Jahoda, 1970; Lancaster, 1968; Veyette, 1961).

It is worth noting that in some instructions, the relevance of keywords is becoming apparent, although the focus is more on the dissemination of articles on the Internet than on the efficiency of information retrieval. In light of the aforementioned considerations, it can be concluded that the criteria proposed by scientific journals to authors for selecting the languages in which they represent the knowledge produced remain unduly focused on formal aspects such as the number of keywords. Furthermore, there is a notable tendency to disregard the

potential of such criteria to facilitate the efficient search for information.

In conclusion, it can be stated that the majority of journals are currently unaware of recent advances in the field of indexing and information retrieval, as well as the value of controlled languages in representing and retrieving knowledge. A subsequent line of inquiry will be to examine the role of controlled languages in the retrieval of scientific publications within the databases where they are stored. This should be based on a comparison between the instructions and how articles are indexed in the databases of journal publishers (manual, automatic, or mixed). In the case of automatic indexing, it is essential to examine the specifications of the tools used and how polysemies and synonymies are treated to neutralize silence and documentary noise. Furthermore, it is crucial to understand the linkage of these journals' databases with reference databases such as WoS and Scopus.

REFERENCES

Ahmad, A., Justo, J. L. B., Feng, C., & Khan, A. A. (2020). The impact of controlled vocabularies on requirements engineering activities: a systematic mapping study. *Applied Sciences*, *10*(21), Article 7749. https://doi.org/10.3390/app10217749

Anderson, J. D., & Perez-Carballo, J. (2001). The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing & Management*, *37*(2), 231. https://doi.org/10.1016/S0306-4573(00)00026-1

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval. The concepts and technology behind search*. Pearson.

Barry, J. M. (2004). The site of origin of the 1918 influenza pandemic and its public health implications. *Journal of Translational Medicine*, *2*(3), 1-4. https://doi.org/10.1186/1479-5876-2-3

Garcia-Alsina, M., & Cobarsí, J. (2022). Controlled vocabularies and information retrieval: 1918 Pandemic's scientific literature as an example. *International Journal of Computer and Information Engineering*, *16*(8), 286-293.

Ghanbarpour, A., & Naderi, H. (2019). A model-based method to improve the quality of ranking in keyword search systems using pseudo-relevance feedback. *Journal of Information Science*, *45*(4), 473-487. https://doi.org/10.1177/0165551518799637

Golub, K. (2021). Automated subject indexing: An overview. *Cataloging & Classification Quarterly*, *59*(8), 702-719. https://doi.org/10.1080/01639374.2021.2012311

Harter, S. P. (1975a). A probabilistic approach to automatic keyword indexing. Part I. On the distribution of specialty words in a technical literature. *Journal of the American Society for Information Science*, *26*(4), 197-206. https://doi.org/10.1002/asi.4630260402

Harter, S. P. (1975b). A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing. *Journal of the American Society for Information Science*, *26*(5), 280-289. https://doi.org/10.1002/asi.4630260504

Hong, J.-Y., Suh, E., & Kim, S.-J. (2009). Context-aware systems: A literature review and classification. *Expert Systems with Applications*, *36*(4), 8509-8522. https://doi.org/10.1016/j.eswa.2008.10.071

Ishida, Y., Shimizu, T., & Yoshikawa, M. (2020). An analysis and comparison of keyword recommendation methods for scientific data. *International Journal on Digital Libraries*, *21*(3), 307-327. https://doi.org/10.1007/s00799-020-00279-3

Jahoda, G. (1970). *Information storage and retrieval systems for individual researchers*. Wiley-Interscience.

Keyser, P. (2012). *Indexing: from thesauri to the Semantic web*. Chandos Publishing.

Knobler, S., Mack, A., Mahmoud, A., & Lemon, S. (2005) "1: The story of influenza." The threat of pandemic influenza: Are we ready? In *Workshop Summary* (pp. 60-61). The National Academies Press.

Kwon, S. (2018). Characteristics of interdisciplinary research in author keywords appearing in Korean journals. *Malaysian Journal of Library & Information Science*, *23*(2), 77-93. https://doi.org/10.22452/mjlis.vol23no2.5

Lancaster, F. W. (1968). *Information retrieval systems: Characteristics, testing, and evaluation*. John Wiley.

Leise, F. (2008). Controlled vocabularies: An introduction. *Indexer*, 26(3). https://doi.org/10.3828/indexer.2008.37

Lu, W., Liu, Z., Huang, Y., Bu, Y., Li, X., & Cheng, Q. (2020). How do authors select keywords? A preliminary study of author keyword selection behavior. *Journal of Informetrics*, *14*(4), Article 101066. https://doi.org/10.1016/j.joi.2020.101066

Veyette, J. H., Jr. (1961). Information retrieval: The general nature of IR and indexing Dewey Decimal System Universal Decimal System. Two new systems regional IR centers related developments. *The American Behavioral Scientist (Pre-1986)*, *4*(10), 15.

White, H. (2013). Examining scientific vocabulary: Mapping controlled vocabularies with free text keywords. *Cataloging & Classification Quarterly*, *51*(6), 655–674. https://doi.org/10.1080/01639374.2013.777004

White, H., Willis, C., & Greenberg, J. (2012). The HIVE impact: Contributing to consistency via automatic indexing. In *Proceedings of the 2012 iConference* (pp. 582-584). Association for Computing Machinery. https://doi.org/10.1145/2132176.213229

World Health Organization. (2015, May). World Health Organization best practices for the naming of new infectious diseases. https://www.who.int/topics/infectious_diseases/naming-new-diseases/en/

Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, *4*(3), 1169-1180.