Lenguajes controlados en la indexación de la literatura científica: el caso de la pandemia de 1918

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

RESUMEN
El interés científico por la pandemia de gripe del 1918 se ha visto impulsado por el surgimiento a principios del siglo XXI de las epidemias de neumonía causadas por virus, y más recientemente el COVID-19 surgido en 2020. Esta comunicación es un estudio exploratorio sobre el uso de los lenguajes controlados en la comunidad científica para localizar e identificar, en las bases de datos de literatura científica, el conocimiento generado y necesitado. La investigación tiene dos objetivos: primero, conocer cuáles son los lenguajes controlados relevantes usados por la comunidad científica para representar el conocimiento producido, y segundo, conocer qué papel juegan los lenguajes controlados en la recuperación de la producción científica. La investigación se focaliza como ejemplo en la producción acerca de la pandemia de 1918, almacenada en dos bases de datos ampliamente utilizadas Web of Science y Scopus, y en los vocabularios controlados del ámbito de ciencias médicas y salud. Tras la identificación de palabras clave para localizar artículos que tratan sobre este tema, se seleccionan las revistas científicas de las que se han recuperado los artículos. A continuación, se toma como base las guías e instrucciones a los autores dadas por las revistas donde se han publicado dichos artículos, con el objetivo de analizar el papel que juegan en dichas revistas las palabras clave y los vocabularios controlados para indexar y

recuperar la producción científica en las bases de datos científicas. Los resultados preliminares muestran el poco uso que las editoriales de las revistas científicas hacen de los vocabularios controlados al no incluirlos, en la mayoría de los casos, en las guías para los autores.

Palabras clave: lenguajes controlados en el ámbito de la salud, recuperación de la información científica, pandemia de 1918, edición científica, autoría científica

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

Cómo citar: Garcia Alsina, M., & Cobarsi-Morales, J. (2024). Lenguajes controlados en la indexación de la literatura científica: el caso de la pandemia de 1918. 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. 115–130). Tallinn, Estonia: Pro-Metrics. DOI: 10.47909/978-9916-9974-8-2.87

INTRODUCCIÓN

Las palabras clave son uno de los instrumentos para la representación del conocimiento. Se emplean en la producción científica para su almacenamiento y posterior recuperación en las bases de datos científicas. Esas palabras se recogen, como metadatos, en las bases de datos en el campo de palabras clave. También están presentes para representar el contenido en el título, resumen, y cuerpo del texto, pudiendo este contenido ser objeto de indexación automática. Se constata el uso habitualmente inadecuado de palabras clave en la indexación de la producción científica referente al caso específico de la pandemia de 1918 (Garcia-Alsina y Cobarsi, 2022; World Health Organization, 2015; Barry, 2004; Knobler et al., 2005). Todo ello lo exponemos más adelante. Teniendo en cuenta este contexto, este trabajo tiene como objetivo estudiar los lenguajes que la comunidad científica prefiere para indexar el conocimiento producido referente a la pandemia de 1918, -erróneamente conocida en lenguaje coloquial-, y a menudo también en la literatura científica, como la "gripe española".

La investigación sobre indexación, tanto automática como manual, junto al papel de los lenguajes controlados y el lenguaje natural, ha sido y es ampliamente objeto de estudio (Ishida et al., 2020; Ghanbarpour y Naderi, 2019; Baeza y Ribeiro, 2011; Hong et al., 2009; Anderson y Perez, 2001; Harter, 1975a,b; Jahoda, 1970; Lancaster, 1968;

Veyette,1960). El uso de palabras clave dadas por el autor y los vocabularios controlados son un área de debate y estudio (White, 2012). La identificación de qué palabras clave utilizar oscila entre la selección libre de palabras por parte del autor, o la extracción automatizada (Ishida, et al. 2020; Lu et al., 2020; Ghanbarpour et Naderi, 2019; Kwon, 2018; Zhang, 2008; Harter, 1975a,b). De manera más específica, los lenguajes controlados (como tesauros, ontologías, taxonomías o listas de encabezamientos) contribuyen a representar el conocimiento velando por la univocidad de los significados al tener en cuenta las polisemias y sinonimias existentes (Keyser, 2012; Leise, 2008).

El uso de lenguajes controlados, tanto en la indexación manual como automática, queda en primer lugar a elección de las editoriales que publican la producción científica y los gestores de bases de datos donde dicha producción se almacena. En segundo lugar, en caso de dejar la elección libre a los autores de sus palabras clave, los mismos autores se enfrentan a diversas posibilidades. Una de ellas es decantarse por la selección de palabras clave sin ser conscientes de la relevancia de estas para que su artículo sea encontrado, y sin usar una estrategia para ello (Lu et al, 2020), con el consecuente perjuicio en la indexación eficiente. Otra opción que tienen los autores es la elección voluntaria de lenguajes controlados para seleccionar las palabras más relevantes (Ishida, et al., 2020).

Sea como fuera, el coste de indexación automática frente a la manual decanta la preferencia por parte de las editoriales a una indexación automática (Zhang et al., 2008), corriendo el riesgo de dejar la elección de palabras clave como un hecho secundario. Del mismo modo hay estudios que constatan el uso de palabras clave creadas por autores que son menos eficientes que las extraídas

automáticamente (White, 2013; White et al., 2012). Otros estudios indican que ambas maneras de indexar (humana y automática) se pueden compaginar extrayendo ventajas de ambas (Anderson y Perez, 2001). Además, actualmente las palabras clave y los lenguajes controlados cobran aún más fuerza para la indexación automatizada, especialmente para su recuperación. No obstante, la automatización aún precisa de un mayor desarrollo y la incorporación de vocabularios controlados (Golub, 2021; Ahmed et al., 2020).

Otro aspecto que se han tenido en cuenta en el estudio de la indexación es el distinto uso de palabras clave entre disciplinas. En este sentido, hay estudios previos que indican tendencia a presentar un menor grado de interdisciplinariedad en el uso de palabras clave por parte de autores según disciplinas específicas (Kwon, 2018). Sobre la indexación de literatura referente a la pandemia de 1918, se constata el uso del término de "*gripe española*" no solo en el ámbito coloquial sino también aplicado a la literatura científica (Garcia y Cobarsi, 2022). El uso de términos geográficos asociados a enfermedades va contra recomendaciones de la WHO (World Health Organization, 2015). Además, los resultados concluyentes de algunos estudios refutan el supuesto origen geográfico de la pandemia de 1918 en España (Barry, 2004). Los vocabularios controlados deben lograr coherencia entre la descripción del contenido y su posterior recuperación, mediante una adecuada integración en la base de datos. Por ello, la necesidad de utilizar diferentes términos (incluyendo "gripe española") para recuperar un solo concepto como "pandemia de 1918" señala un fallo en la indexación en las bases de datos de la literatura científica vinculada a este tema, máxime si tenemos en cuenta el

tratamiento de este término en los lenguajes controlados, tanto los generalistas (*Library of Congress Subject Headings* o UNESCO) como los especializados en el ámbito de la salud (MeSH) o humanidades y ciencias sociales (HASSET), por poner algún ejemplo. Una exploración en el Registro Basado en Tesauros, Ontologías y Clasificación (BARTOC) apunta a términos concretos vinculados a pandemia o influenza, excluyendo el uso del término "gripe española".

En definitiva, este trabajo aborda las instrucciones que las revistas dan a los autores sobre el uso de las palabras claves, y por tanto, el marco con el que se encuentran para que su trabajo sea indexado. Esta fase de la investigación parte de la siguiente pregunta: ¿Cuáles son los criterios propuestos por parte de las revistas científicas a los autores para seleccionar los lenguajes con los que etiquetan el conocimiento producido?

METODOLOGIA

La investigación toma como base los artículos producidos entre 2000 y 2019 sobre la pandemia de 1918, y que están indexados en dos bases de datos: Web of Science y Scopus. La elección de los años viene motivada por el interés surgido a partir de 2003 en este tema a raíz del inicio de la epidemia de SARS que motivó un incremento de investigaciones y previo a la pandemia de COVID-19.

Para localizar revistas sobre las que hacer el trabajo de campo, se buscaron los artículos empleando cuatro palabras clave en inglés: "Spanish influenza", "Spanish flu", "1918 influenza" y "1918 flu". Los términos seleccionados contemplaron sinonimias en cuanto a la enfermedad propiamente dicha y la variedad en el término que acompaña a la enfermedad (país y año). Con estas búsquedas

hemos obtenido un total de 70 artículos, publicados en 61 revistas. De estas revistas hemos localizado su página web, desde donde podíamos obtener la guía de publicación para los autores. Tras observar las webs, excluimos del estudio algunas revistas siguiendo los siguientes criterios: revistas que han publicado artículos divulgativos o de debate; las que se editan en un idioma desconocido para los autores de este estudio (coreano, islandés, noruego y sueco), al no poder identificar cuáles eran las instrucciones a los autores; y, por último, las que ya no se editan a fecha de hoy y, por lo tanto, carecen de acceso a las instrucciones que en ese momento tenían los autores. En total hemos trabajado con un listado de 49 revistas.

La información extraída de las instrucciones a autores es la siguiente:

- El ámbito al que pertenece la revista: ciencias de la salud, ciencias experimentales, ingeniería informática, ciencias sociales, humanidades e interdisciplinar.
- La existencia de instrucciones a los autores sobre cómo seleccionar las palabras clave.
- La especificación de si un vocabulario controlado se debe usar, o si los términos a emplear son de libre creación.
- El vocabulario que el autor debe usar, si es el caso.
- La indicación vinculada a SEO, si es el caso.

RESULTADOS

El análisis del contenido de las instrucciones para autores en las webs de las revistas evidencia un predominio de revistas de ciencias de la salud, seguido de las ciencias sociales y humanidades (Tabla 1).

Tabla 1. Ámbito temático de las revistas (Fuente: elaboración propia).

ÁMBITO TEMÁTICO	N° DE REVISTAS	%
Ciencias de la salud	29	59,18
Humanidades	8	16,32
Ciencias sociales	6	12,24
Ciencias experimentales	4	8,16
Ingeniería informática	1	2,04
Interdisciplinar	1	2,04

Del total de las revistas examinadas, una fracción considerable de ellas (59,18%) ofrece instrucciones a los autores. La mayoría de las revistas pertenecen al ámbito de las ciencias de la salud (68,97%), junto con humanidades y ciencias sociales, ofrecen instrucciones a los autores sobre palabras clave (10,34%).

Tabla 2. Ámbitos de las revistas con instrucciones a autores (Fuente: elaboración propia).

ÁMBITO	% REVISTAS CON INSTRUCCIONES (SOBRE EL TOTAL DE REVISTAS CON INSTRUCCIONES)	N° DE REVISTAS
Ciencias de la salud	68,97	20
Ciencias sociales	10,34	3
Humanidades	10,34	3
Ciencias experimentales	6,90	2

Table 2. *Continued*

ÁMBITO	% REVISTAS CON INSTRUCCIONES (SOBRE EL TOTAL DE REVISTAS CON INSTRUCCIONES)	Nº DE REVISTAS
Ingeniería informática	3,45	1
Interdisciplinar	0	0

Las revistas que carecen en absoluto de indicaciones respecto a las palabras clave en sus guías a los autores para publicar son aún un número elevado (40,81%), y se deduce de ello falta de valoración de las palabras clave por parte de estas revistas.

Tabla 3. Indicaciones sobre palabras clave (Fuente: Autor).

INSTRUCCIONES SOBRE PALABRAS CLAVE	Nº DE REVISTAS	% DE REVISTAS
Revistas sin instrucciones	20	40,81%
Revistas con instrucciones	29	59,18%

En el caso de las revistas que dan indicaciones, de manera mayoritaria las palabras clave se dejan a libre elección del autor, siendo minoría las revistas que proponen el uso de algún vocabulario controlado. De este modo, la selección de palabras clave y su correspondiente indexación queda en manos de los autores y en riesgo de que el contenido sea difícil de recuperar en las búsquedas. En la Tabla 4 se resume esta faceta.

Tabla 4. Indicaciones sobre palabras clave (Fuente: elaboración propia).

INSTRUCCIONES SOBRE PALABRAS CLAVE	Nº DE REVISTAS	% DE REVISTAS
Palabras clave de libre elección	20	69%
Palabras clave mediante lenguajes controlados	9	31%

En el caso de palabras clave a libre elección del autor, una directriz habitual se refiere a establecer un número mínimo y/o máximo de palabras clave (lo hacen 23 de las 29 revistas que ofrecen instrucciones a los autores). No acostumbran a apuntarse otras directrices, excepto en algunas de ellas para aconsejar el uso de términos dirigidos a una mayor divulgación de los artículos.

Por lo que respecta a las revistas que indican el uso concreto de un vocabulario controlado, destacan tres lenguajes, dos lenguajes son del ámbito de la Salud y uno de Ciencias Sociales. Los dos lenguajes del ámbito de la salud son el MeSH (Medical Subject Headings) y el CINAHL (Cumulative Index to Nursing and Allied Health). El del ámbito de ciencias sociales es el JEL, un sistema de clasificación creado por el *Journal of Economic Literature. Este* es un método estándar para clasificar la literatura científica del campo de economía, tal como se indica en la guía para autores (Tabla 5)

Si además de lo descrito consideramos el total de revistas estudiadas (49) y el número de revistas que emplean lenguajes controlados (9), constatamos el bajo uso que las revistas fomentan para representar e indexar

Tabla 5. Uso de lenguajes controlados en las revistas con instrucciones (Fuente: elaboración propia).

VOCABULARIO	ÁMBITO	Nº DE REVISTAS	% DE REVISTAS (EN RELACIÓN CON LAS QUE DAN INSTRUCCIONES)	% DE REVISTAS (EN RELACIÓN CON LA TOTALIDAD DE LAS EXAMINADAS)
MeSH	Ciencias de la Salud	6	66,66%	12,24%
MeSH y CINAHL	Ciencias de la Salud	2	22,22%	4,08%
JEL	Ciencias sociales	1	11,11%	2,04%

el conocimiento, puesto que solo el 18,36% impulsan el uso de los lenguajes controlados. Por otro lado, emerge en algunas revistas (12,24%) la necesidad de concienciar a los autores de la relevancia de las palabras clave no solo en la sección de palabras clave, sino también en el título, resumen y en el propio texto del artículo. Todas las recomendaciones están orientadas a trabajar el SEO (Search Engine Optimization) para que los artículos sean localizados en internet, ya sea en Google Scholar o en otros repositorios abiertos. Ninguna de las instrucciones se refiere a la indexación automática de las bases de datos de las editoriales. Asimismo, es de resaltar el foco que se pone en el SEO antes que en la eficacia y calidad de la recuperación de la información, para eliminar el ruido y el silencio documental. Analizando las instrucciones, se observa que las revistas que valoran y hacen hincapié en el SEO, ven a éste desde la perspectiva de la difusión de la producción de los autores, y por ende de la propia revista. La recuperación de información de manera más relevante y exhaustiva no está detrás de la importancia de las palabras clave que se da en estas revistas y sus instrucciones.

De lo examinado, queda claro que las palabras clave, y en particular la optimización de estas en orden a la indexación y recuperación de los artículos, no son ni mucho menos un requerimiento prioritario a los autores por parte de las revistas. Ello es todavía más remarcable si se comparan con requerimientos mucho más frecuentes y explícitos en las instrucciones a autores por parte de las revistas, tales como formateo de referencias bibliográficas, antiplagio, etc.

Hasta aquí nuestro análisis sobre las instrucciones de las revistas. Como limitación podemos señalar que únicamente se han consultado las instrucciones publicadas en la

web de libre acceso de las respectivas revistas. En este sentido, no se han utilizado los formularios y aplicativos que muchas publicaciones tienen para recoger los envíos, susceptibles de contener instrucciones adicionales incrustadas en su interfaz.

DISCUSION

Los resultados de esta investigación preliminar evidencian que las editoriales de las revistas científicas escasamente emplean los lenguajes controlados para representar el conocimiento, indexarlo y recuperarlo de manera eficiente y relevante, al no incluirlos en las instrucciones a los autores. Por tanto, no son conscientes del potencial de estos lenguajes para neutralizar los ruidos y los silencios documentales. Este uso parece estar al margen de los requerimientos de los sistemas de recuperación de la información y los instrumentos de indexación, automática y manual, identificados por investigadores de este campo, tal como apuntan algunos de los innumerables estudios en la materia (Ishida et al., 2020; Ghanbarpour y Naderi, 2019; Baeza y Ribeiro, 2011; Honget al., 2009; Anderson y Perez, 2001; Harter, 1975a,b; Jahoda, 1970; Lancaster, 1968; Veyette,1960).

Es de subrayar que en algunas instrucciones se está manifestando la relevancia de las palabras clave, aunque más enfocada a la difusión de los artículos en internet que a la eficiencia en la recuperación de la información. Por tanto, respondiendo a nuestra pregunta de investigación, sobre los criterios propuestos por parte de las revistas científicas a los autores, para que estos seleccionen los lenguajes con los que representan el conocimiento producido, podemos concluir que están aún demasiado focalizadas a aspectos formales como el número de palabras clave, y aún

hay tendencia amplia a ignorar el potencial que tienen para facilitar la búsqueda eficiente de la información.

En conclusión, en estos momentos, las revistas están mayoritariamente al margen de los avances de investigación en materia de indexación y recuperación de la información, y del valor que los lenguajes controlados tienen para representar y recuperar el conocimiento. Siguiendo con esta investigación, una futura línea de trabajo será contemplar qué papel juegan los lenguajes controlados en la recuperación de la producción científica en las bases de datos donde se almacenan. Se debería partir de la comparativa entre las instrucciones y como se indexan los artículos en las bases de datos de las editoriales de las revistas (manual o automática o mixta). En caso de ser automática, se debería ver las especificaciones de las herramientas y cómo se tratan las polisemias y las sinonimias para neutralizar silencio y ruido documental. También se debería conocer el vínculo de las bases de datos de estas revistas con bases de datos referentes como Web of Science y Scopus.

REFERENCIAS

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.