

XXIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXIV ENANCIB

ISSN 2177-3688

GT 2 – Organização e Representação do Conhecimento

EL CONTROL DE VOCABULARIO COMO HERRAMIENTA PARA MAPEAR INTERESES DE INVESTIGACIÓN: EL CASO DE LAS CIENCIAS AGRARIAS EN URUGUAY

VOCABULARY CONTROL AS A TOOL FOR MAPPING RESEARCH INTERESTS: THE CASE OF AGRICULTURAL SCIENCES IN URUGUAY

Lucía Simón – Universidad de la República (UdelaR)

Mario Guido Barité – Universidad de la República (UdelaR)

Modalidade: Resumo Expandido

Resumen: Mejorar la calidad de vida de las personas, es fundamental en lo que hace a las cuestiones del desarrollo. Por esto, conocer los intereses de investigación que colaboran con dichas mejoras resulta relevante. Se propone explorar cuáles son estos intereses en ciencias agrarias en Uruguay y en un futuro, estudiar a la solución de qué problemáticas del desarrollo pueden contribuir a resolver. Para alcanzar este objetivo, en este trabajo parcial, se identifican cuáles son los intereses de investigación en ciencias agrarias, mediante técnicas de control de vocabulario. Existen diversos estudios que indagan cuáles son los temas de investigación en algún ámbito determinado, y en su extensa mayoría, utilizan los CV de quienes investigan como fuente de datos. Mediante el uso de herramientas estadísticas (bibliométricas) se obtienen resultados que pretenden dar cuenta de dichas temáticas. Este trabajo, aporta una aplicación novedosa al uso de una herramienta utilizada en las ciencias de la información: el control de vocabulario (mediante técnicas de indización). Se parte de la premisa, que el control de vocabulario es una herramienta más robusta, con base en dos supuestos: casi ninguna herramienta reconoce sintagmas a la hora de extraer los términos de los textos y, la vinculación de temas de investigación con problemas que éstos contribuyen a resolver es un enfoque novedoso en este tipo de estudios. Finalmente, se presentan las temáticas de investigación de las ciencias agrarias en Uruguay y su posible aporte a la solución de problemas del desarrollo.

Palavras-chave: control de vocabulario; intereses de investigación; problemas del desarrollo

Abstract: The improvement of people's quality of life is essential when it comes to development issues. For this reason, to identify the research interests that contribute to these improvements is relevant. This research aims to explore which interests are observed in Agricultural Sciences in Uruguay. In the future, it aims to study which kind of development problems they may contribute to solve. For that, in this abstract, research interests are identified in Agricultural Sciences through vocabulary control techniques. There are various studies that investigate which are the research topics in each field, and in their vast majority, they use the CVs of those who research as a source of data. To achieve this goal, it is used statistical tools (bibliometric). This work provides a novel application to the use of a tool used

in Information Sciences: vocabulary control (through indexing techniques). The starting point is the assumption that vocabulary control is a more robust tool, based on two features: almost no tool recognizes phrases when extracting terms from texts, and the relation among research topics and problems that they contribute to solve is a novel approach in this type of study. Finally, the research topics of Agricultural Sciences in Uruguay and their possible contribution to the solution of development problems are presented.

Keywords: vocabulary control; research interests; maintenance problems

1 INTRODUCCIÓN

La indagación sobre qué se investiga en un dominio determinado no es una novedad. Diversos estudios a lo largo de los años han aportado evidencia en este sentido, sobre todo a través de estudios métricos de la información, facilitados por el acceso a grandes bases de datos (como ser WoS, Scopus, entre otras), o por la disponibilización de currículums vitae estandarizados del cuerpo de quienes investigan en determinado lugar. A modo ilustrativo, en la tesis de Rafael Antunes dos Santos (2015) se presenta una síntesis de diversos estudios que, a través del análisis de coocurrencias, han mapeado distintas estructuras de conocimiento a lo largo del tiempo.

Este tipo de estudios dan cuenta de diversas cuestiones: áreas de investigación emergentes, redes de interacciones, colegios invisibles, frentes de investigación, mapeo de campos científicos, fuentes de investigación, mediante el uso de palabras de los títulos, palabras clave (proporcionadas por lo autores o propias de sistemas de indización), indización automática, descriptores de asunto (como es el caso de MESH), tags, o combinaciones de éstas (palabras clave + palabras del título + palabras de los resúmenes) (Santos, 2015). Sin embargo, el uso de herramientas propias de la organización del conocimiento (como es la utilización de vocabularios controlados) para realizar este tipo de estudios, ha sido escasamente utilizado.

Más allá de contar con esta información: ¿cuáles son los temas sobre los que se investiga?, que también podría traducirse como: ¿cuáles son las capacidades (nacionales, locales, dependerá del caso) existentes de investigación que pueden colaborar en la solución de problemas?, los países subdesarrollados aún mantienen vigente la premisa de que estas regiones, se caracterizan por la baja incorporación de conocimiento endógeno, o lo que es lo mismo, por la subutilización de sus capacidades de investigación para la resolución de problemas (Fajnzylber, 1992). Este trabajo, parte del supuesto que uno de los motivos por los que esas capacidades ya mapeadas no son utilizadas (o son subutilizadas) es que su

representación, recolectada con técnicas como las descritas en el párrafo anterior, ofrece aproximaciones más generales (e imprecisas, desde el punto de vista terminológico) de las que pueden realizarse con el uso de técnicas de control de vocabulario. En función de esto, se propone evidenciar cómo las técnicas de control de vocabulario aportan precisión en la tarea de mapear estructuras de conocimiento científicas, a través de un estudio de caso, centrado en las ciencias agrarias en Uruguay.

Este trabajo presenta, en primer lugar y de forma resumida, antecedentes referidos a estudios que han mapeado temas de investigación y se presentan algunas primeras consideraciones sobre la eventual ventaja de utilizar técnicas como el control de vocabulario para realizar esta tarea. A continuación, se explicitan los pasos que se siguieron para cumplir con el objetivo propuesto y se presentan los resultados obtenidos. Finalmente, se esbozan algunas conclusiones primarias referidas a la utilidad y robustez del control de vocabulario (CV) como herramienta para mapear intereses de investigación.

2 ANTECEDENTES

A nivel nacional e internacional se encontraron numerosos trabajos cuyo objetivo ha sido mapear las estructuras temáticas de conocimiento de diversos dominios. En algunos casos, los estudios se centran en un área del saber, en otros, en las temáticas presentes en un dominio determinado, como lo es el campo de los Estudios Sociales de la Ciencia y la tecnología, entre otros (Jacobs, 2012; Herrera-Miranda *et al.*, 2013; Matti; Dutta, 2013; Santos, 2015; Satsumi; Berdón; Grimaldi, 2024; Levin; Davyt; Invernizzi, 2024). Estos antecedentes tienen dos diferencias sustantivas con el trabajo propuesto: la primera, vinculada con la fuente de datos utilizada para realizar la investigación, y la segunda, referida a la técnica empleada para alcanzar el objetivo de mapear las estructuras temáticas de conocimiento de un dominio determinado.

Referido a la primer diferencia mencionada: las fuentes de datos, se destaca que los estudios precedentes utilizan (en general), dos fuentes de datos: i) Bases de datos que recogen la producción científica de quienes investigan (como lo son la WoS, Scopus, Dimensions) y ii) CV. Esta última fuente, es utilizada cuando existe un formato único (o principal, como lo es el Cvuy en Uruguay, o el CVLattes en Brasil). Con cualquiera de estas dos fuentes, y en relación a la segunda diferencia planteada, los estudios realizan distintos tipos de operaciones (sobre todo basadas en el análisis de coocurrencias) que les permiten determinar, a grandes rasgos,

cuáles son las principales tendencias temáticas. Los análisis referidos incluyen el análisis de títulos, de palabras clave y de resúmenes. Como se mencionó, el resultado determina a grandes rasgos las principales tendencias temáticas de investigación. Por lo general, estas son plasmadas tal y como aparecen sin realizar ninguna operación sobre ellas.

3 EL CONTROL DE VOCABULARIO COMO HERRAMIENTA PARA MAPEAR INTERESES DE INVESTIGACIÓN

El control de vocabulario puede considerarse como una combinación de técnicas y procedimientos que se aplican al lenguaje de forma consistente para establecer expresiones estandarizadas (Barité *et al.*, 2015). En ese sentido, estas expresiones aspiran a ser más adecuadas para la representación y la recuperación de la información; además, favorecen la comunicación. Esto último, en el entendido de que el lenguaje natural se expresa de variadas formas y puede causar problemas o dificultades a la hora de la comunicación. El control de vocabulario, entonces, se presenta como una buena herramienta para desambiguar las inconsistencias del lenguaje natural Hjørland (2008).

Como se mencionó, el mapeo de intereses de investigación (o las distintas denominaciones que se utilizan) es un tema sobre el que existe un importante número de trabajos realizados. No obstante, las técnicas empleadas para mapearlos responden a los denominados estudios métricos de la información, donde no están presentes operaciones de control de vocabulario (al menos en la mayoría de los casos). En este marco, los escasos estudios que refieren específicamente al uso de técnicas de control de vocabulario, utilizan esta herramienta en combinación con técnicas bibliométricas (como las mencionadas más arriba), es decir, su uso es marginal y responde a estandarizar algunas de las expresiones recogidas.

Por lo antes mencionado, se entiende que proponer al control de vocabulario como una herramienta teórico-metodológica que permite (de forma precisa) determinar los intereses de investigación en cualquier ámbito especializado del conocimiento, constituye una novedad y un aporte, tanto a los estudios vinculados con el comportamiento científico, como a la ciencia de la información.

4 METODOLOGÍA

Para alcanzar los objetivos propuestos, se diseñó un estudio observacional (Kisch, 2004) de corte cualitativo mediante la aplicación de técnicas de control de vocabulario. La fuente de datos utilizada, fue una encuesta realizada entre agosto y diciembre de 2022, destinada a las personas que investigan en Uruguay. La misma, tuvo lugar en el marco del núcleo interdisciplinario: CiTINDe - Ciencia, Tecnología e Innovación para un Nuevo Desarrollo y se titula: “Opinión de quienes investigan acerca de diversos aspectos de sus procesos de trabajo y cómo mejorarlo”. La población total de la encuesta suma un total de 3482 personas, de las que 1112 respondieron (31.9%).

Para este trabajo, se seleccionaron aquellas personas que indicaron su área de adscripción como ciencias agrarias cuya población total representa el 15.5% del total (539 personas). De esas 539 personas, 155 respondieron efectivamente esta encuesta. Si bien el número puede parecer despreciable, un análisis detallado de la validez de la muestra obtenida puede consultarse en Concitinde (2023). Con foco en el trabajo que aquí se presenta, se optó por analizar dos ítems abiertos (y sus palabras clave asociadas, proporcionadas por quien completa la encuesta) que se presentaron en un pequeño bloque titulado: “Orientación del trabajo de investigación”. En este bloque, analizamos el primer ítem y sus palabras clave: “indique en hasta 100 palabras sus principales intereses de investigación”. El segundo ítem, está siendo analizado (este, vincula los intereses de investigación con la posibilidad de resolver problemas sociales y/o asociados a la producción).

De las respuestas obtenidas, y para poder analizar cuáles son los intereses de investigación en las ciencias agrarias en Uruguay, se chequeó que, al menos, las personas que contestaron la encuesta, hayan respondido a la pregunta sobre sus intereses de investigación. En este caso, el total de respuestas es de 147. Este es el número de respuestas que fue analizado y sobre el cual se presentan los primeros resultados.

Para determinar cuáles son los intereses de investigación, se estableció el siguiente procedimiento en 4 fases:

1. Exploración primaria: Análisis de frecuencias.

- i) Se elaboró un listado con frecuencias simples de las palabras presentes en el campo abierto del ítem a analizar y de las palabras clave asignadas. ii) Se eliminaron las palabras no

significativas. iii) Se normalizaron los sinónimos; plurales y singulares. iv) Se corrigieron errores ortográficos.

Esta operación se llevó adelante para evitar posibles pérdidas de información (se utilizó para comparar con el listado final, resultado de la segunda fase).

2. Marcado y selección de términos candidatos.

i) Implicó la lectura de cada respuesta y de las palabras clave asignadas. ii) Marcado de términos candidatos. iii) Fijación de los mismos en función de dos criterios: existencia del término en el tesoro AGROVOC, o alta garantía literaria (validada con Google Scholar). iv) Creación de una columna donde se consignaron los términos candidatos a representar intereses de investigación.

3. Normalización de términos candidatos.

Se realizaron las siguientes operaciones sobre cada término, lo que determinó cambios, exclusiones y alguna incorporación: restricción (en caso de ser necesario y de que la definición existiera en el Tesoro); desambiguación (control de homónimos, términos polisémicos, sinónimos y variantes); jerarquización y relación. Finalmente, se revisó el listado completo obtenido.

4. Revisión del listado final.

5 PRIMEROS RESULTADOS Y DISCUSIÓN

Se presentan y discuten los primeros resultados obtenidos siguiendo los pasos detallados en la metodología.

La tarea de identificar cuáles son los presuntos intereses de investigación, siguió las fases indicadas en el apartado metodológico para el que se precisaron las siguientes acciones:

1. En un primer momento, se confeccionaron tres listados (que se plasmaron en una única plantilla Excell), en función de la procedencia del término candidato, a esto le llamamos: “Tipo de lenguaje” y la tipología es la siguiente:

i) Control de vocabulario. Incluye los términos que son candidatos como resultado de operaciones de control de vocabulario. Estas operaciones implican dos acciones: buscar el presunto término (no tiene por qué ser exacto) en AGROVOC, si aparece, se consigna tal y como se registra en el tesoro. En caso de que el término no se encuentre en el tesoro AGROVOC pero se crea que es pertinente incluirlo, se lo buscó entrecomillado en el buscador Google Scholar. Se determinó que si hay más

de 500 documentos asociados al término, la garantía literaria será fuerte, por lo que se incorporará el término.

ii) Palabras clave. Incluye las palabras clave asignadas por quienes completaron la encuesta.

iii) Lenguaje natural. Incluye todos los términos (depurados) que se elaboraron en la fase 1 de la metodología (frecuencia simple).

2. Se marcaron las palabras clave proporcionadas por las personas que investigan que no presentaron coincidencias con los términos seleccionados manualmente (con técnicas de control de vocabulario), pero que estaban recogidas en el tesoro AGROVOC o que presentaban niveles no despreciables de garantía literaria.

3. Se marcaron los términos derivados de la frecuencia simple de las palabras presentes en los ítems de análisis de la encuesta: Se marcaron los términos coincidentes con términos seleccionados manualmente (con técnicas de control de vocabulario).

En los todos los casos, se procedió a: i) controlar los términos duplicados; ii) corregir eventuales errores tipográficos; iii) normalizar el uso de plurales y singulares; iv) controlar los sinónimos y v) eliminar las palabras no significativas.

Tabla 1 – Cantidad de términos seleccionados como posibles intereses de investigación según tipo de lenguaje

Tipo de lenguaje	Cantidad de términos	Porcentaje
Control de vocabulario – CV (manual + AGROVOC o garantía literaria)	240	43.9
Palabras clave - PC	77	14.1
Lenguaje natural (frecuencia simple) - LN	37	6.8
CV-PC-LN	18	3.2
CV-PC	37	6.8
CV-LN	129	23.6
PC-LN	8	1.5
Total	546	99.9

Fonte: Dados de investigación.

Lo primero que se evidencia, es que cerca de la mitad de los términos corresponde solo a un tipo de lenguaje específico: selección de términos con operaciones de control de vocabulario (el 43.9%), mientras que las palabras clave proporcionadas por quienes

investigan, representan casi un tercio del total (14.1%). El total de palabras que aparecen, producto del conteo simple, es muy bajo, representando el 6.8% del total.

Por su parte, los términos marcados a través de técnicas de control de vocabulario que coinciden con los proporcionados por los autores, alcanzan un porcentaje también bajo, pero implican que esos términos, se habrán de fijar para el mapeo de los intereses de investigación. Asimismo, los términos controlados y los términos que aportó la frecuencia simple, son los que alcanzan un porcentaje mayor, pero este resultado no es significativo, ya que, como, es de esperar, los términos que derivan de la frecuencia de palabras, se encuentren en el resto de los lenguajes al tratarse de términos no compuestos. Finalmente, los términos proporcionados por los autores que coinciden con la frecuencia simple, son 8 (representando el 1.5% del total). Además de representar un magro número, son términos generales, que, en todos los casos, tienen un término sintagmático (más específico) asociado.

A continuación, se presentan los resultados obtenidos con relación a la presencia de los términos seleccionados (y normalizados) en el vocabulario controlado AGROVOC y su garantía literaria. Esto último, también se detalla para aquellos términos que no estaban presentes en el lenguaje controlado y se procurará defender su inclusión (o no) en la categorización final, en función de la garantía literaria presente.

En lo que respecta a la garantía literaria, se establecieron dos categorías para determinar su nivel, a saber: 1. Nivel medio de garantía literaria (menos de 500 documentos recuperados de Google Scholar, mediante la búsqueda exacta del término); 2. Nivel alto de garantía literaria (más de 500 documentos recuperados de Google Scholar con el mismo procedimiento).

Tabla 2 – Cantidad de términos presentes en vocabulario controlado y presunta garantía literaria

Tipo de vocabulario	AGROVOC	NO AGROVOC (sólo GL)
Cantidad de términos	442 (81%)	104 (19%)

Fonte: Datos de investigación.

Tabla 3 – Nivel de garantía literaria de los términos presentes en AGROVOC (442)

Nivel de garantía literaria	N.º de términos presentes (AGROVOC)	%
1 (nivel medio)	33	7.5
2 (nivel alto)	409	92.5

TOTAL	442	100
-------	-----	-----

Fonte: Dados de investigação.

Tabla 4 – Nivel de garantía literaria de los términos que no se encontraron en AGROVOC

Nivel de garantía literaria	N.º de términos solo con GL	%
1 (nivel medio)	25	24
2 (nivel alto)	79	75.9
TOTAL	104	99.9

Fonte: Datos de investigación.

La tabla 3, evidencia que del total de términos candidatos, una amplísima mayoría tiene respaldo en el vocabulario controlado AGROVOC (81% del total de los términos). Ahora bien, es necesario validar el resto de los términos, y “probar”, que los que se encuentran en el vocabulario controlado, poseen amplio respaldo en la literatura. Sobre este punto, la tabla 4, muestra que el 92.5% de los términos incluidos presentes en AGROVOC, poseen una garantía literaria alta.

La tabla 4, muestra que de un total de 104 términos que no están presentes en AGROVOC, un porcentaje menor que los de la tabla anterior (pero igualmente significativo), tiene alto presencia de garantía literaria (75.9%). Dicho esto, cabe realizar una aclaración para todos los casos: cuando el término es demasiado genérico, y claramente escapa exclusivamente al campo de las ciencias agrarias, el resultado de la garantía literaria “se dispara”: esto es razonable, ya que se trata de términos que son abordados desde distintos dominios). El listado final de los términos, puede descargarse aquí: <https://cloud.csic.edu.uy/index.php/s/KTaFIPv29KKwBT2>

6 CONCLUSIONES PRIMARIAS

Este trabajo pretendió mostrar todos los pasos seguidos para identificar cuáles son los intereses de investigación en ciencias agrarias en Uruguay. El número de intereses (que podrían ser homologables con temáticas) arrojó un total de 546. Si bien, para este trabajo no se realizó ninguna agrupación jerárquica (la misma está siendo analizada en este momento, por lo que se aspira a presentar estos resultados en ENANCIB) el nivel de detalle que presentan estos términos, es mucho mayor que los encontrados en la bibliografía que mapea temas de investigación.

El uso de técnicas de control de vocabulario y herramientas derivadas del mismo (en este caso, el vocabulario controlado AGROVOC) demuestra ser robusto para estos fines. Se pretende buscar el modo de agrupar estos intereses en categorías más amplias, pero que puedan evidenciarse cada uno de ellos, pues se considera importante contar con esta información a la hora del diseño de instrumentos relacionados con la política en ciencia y tecnología del país.

REFERENCIAS

BARITÉ, M. *et al.* **Diccionario de Organización del Conocimiento: Clasificación, Indización, Terminología**. 6. ed. Montevideo: CSIC, 2015.

HERRERA-MIRANDA, I.; LICEA DE ARENAS, J.; GÓMEZ-HERNÁNDEZ, J. A. Publicaciones Periódicas en Biblioteconomía, Bibliotecología, Ciencias de la Información y Documentación en México: Tendencias Temáticas, Productividad y Redes de Coautoría: 1956-2006. **Revista Interamericana de Bibliotecología**, v. 36, n. 2, p. 97-108, 2013.

HJØRLAND, B. What is Knowledge Organization (KO)? **Knowledge Organization**, v. 35, n. 2-3, p. 86-101, 2008.

JACOBS, N. Co-term network analysis as a means of describing the information landscapes of knowledge communities across sectors. **Journal of Documentation**, v. 58, n. 5, p. 548-562, 2012.

KISH, L. **Statistical Design for Research**. New Jersey: Wiley-Interscience, 2004.

LEVIN, L.; DAVYT, A.; INVERNIZZI, N. Dinámicas de las agendas de investigación en las Jornadas ESOCITE, 1996-2018. **Sociologias**, Porto Alegre, v. 26, p. e-soc132468, 2024, p. 1-37.

MAITI, D.; DUTTA, B. Comparative Study Between Words in Titles and Keywords of Some Articles on Knowledge Organisation. **DESIDOC Journal of Library & Information Technology**, v. 33, n. 6, p. 498-508, 2013.

MONTILLA PEÑA, L. J. Análisis de la producción científica de los artículos de la Revista Zootecnia Tropical del Instituto Nacional de Investigaciones Agrícolas (2006-2013). **Biblios**, v. 65, p. 1-14, 2016.

SANTOS, R. A. Análise de coocorrência de palavras na pesquisa brasileira em HIV/AIDS indexada na Web Of Science no período 1993-2013. 2015. Tese (Mestrado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2015.

XXIV Encontro Nacional de Pesquisa em Ciência da Informação – XXIV ENANCIB
Vitória-ES – 04 a 08 de novembro de 2024

SATSUMI, J.; BERDÓN, V.; GRIMALDO, C. Relación entre los Programas Nacional Estratégicos de CONHACYT y los Objetivos del Desarrollo Sostenible: Análisis Utilizando ATLAS.ti. **Ciencia Latina Revista Multidisciplinar**, v. 8, n. 3, p. 1868-1888, 2024.