VISUALIZACIÓN Y MÉTRICAS PARA EL ANÁLISIS DEL CICLO DE VIDA DE LA INDIZACIÓN: ESTUDIO DE CASO DE 25 AÑOS DE *KEYWORDS* EN EL DIARIO *EL PAÍS* (2000-2024) Y PROTOTIPO DE APLICACIÓN *WEB*

TOMÁS SAORÍN*
JUAN-ANTONIO PASTOR-SÁNCHEZ**
ISIDORO GIL-LEIVA***

Resumen: *Mientras que la indización condensa grandes volúmenes de información en términos representativos que facilitan su organización y recuperación, la visualización de información permite identificar y representar gráficamente patrones y estructuras relevantes en los datos. En este trabajo se presenta el diseño, desarrollo y validación de una aplicación* web *para la explotación y visualización dinámica de un corpus de indización compuesto por más de 73.000 términos diferentes y generado durante 25 años por el periódico español* El País*. Se ha adoptado un enfoque metodológico de carácter cuantitativo-computacional, que integra el procesamiento estadístico de los datos, el modelado semántico y la representación visual interactiva. Los resultados evidencian una herramienta robusta conformada por numerosas métricas e indicadores que hacen visible, comprensible e interpretable la indización con un potencial de aplicación en diversas áreas académicas, y fácilmente adaptable a otros conjuntos de datos provenientes de otros periódicos.*

Palabras clave: *Indización; Medios de comunicación;* El País*; Visualización de la información; Modelado de métricas e indicadores.*

Abstract: *While indexing condenses large volumes of information into representative terms that facilitate their organization and retrieval, information visualization enables the identification and graphical representation of relevant patterns and structures within the data. This study presents the design, development, and validation of a web application for the exploration and dynamic visualization of an indexing corpus comprising more than 73.000 distinct terms, generated over 25 years by the Spanish newspaper* El País*. A quantitative-computational methodological approach has been adopted, integrating statistical data processing, semantic modelling, and interactive visual representation. The results demonstrate a robust tool composed of numerous metrics and indicators that make the indexing visible, comprehensible, and interpretable, with potential applications across various academic fields and easy adaptability to other datasets from different newspapers.*

Keywords: *Indexing; Media;* El País*; Information visualization; Metric and indicator modelling.*

* Universidad de Murcia – España. Email: tsp@um.es. ORCID: https://orcid.org/0000-0001-9448-0866.
** Universidad de Murcia – España. Email: pastor@um.es. ORCID: https://orcid.org/0000-0002-1677-1059.
*** Universidad de Murcia – España. Email: isgil@um.es. ORCID: https://orcid.org/0000-0002-7175-3099.

INTRODUCCIÓN

La indización con palabras clave (*keywords*) o descriptores es un recurso que ha sido ampliamente utilizado para potenciar la recuperación de información, tanto en el campo de las bases de datos de documentación científica, como en otros sistemas específicos como las bases de datos de legislación y jurisprudencia y las noticias y archivo de medios de comunicación.

El proceso de indización ha sido definido por la norma ISO 5963-1985 (International… 1985) como «el acto de describir o identificar un documento en términos de su contenido temático». La indización puede desarrollarse en cuatro entornos: (1) la indización del autor, frecuente en publicaciones científicas, donde se solicita a los investigadores que proporcionen palabras clave representativas; (2) la indización profesional, realizada por documentalistas que emplean tanto lenguaje natural como controlado; (3) la indización social, llevada a cabo de manera colaborativa por comunidades de usuarios; y (4) la indización automática, desarrollada por sistemas informáticos capaces de extraer tanto términos textuales como descriptores normalizados. Estos entornos pueden funcionar combinados en diferentes proporciones y flujos, según las necesidades de cada sistema o entorno, y apoyan funciones esenciales en los procesos de almacenamiento y recuperación, como se esquematiza en la siguiente figura.

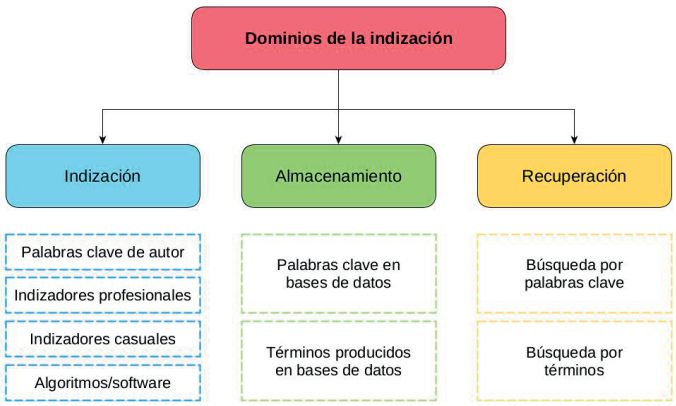


Fig. 1
Dominios de la indización
Fuente: Gil-Leiva 2025, p. 519

Tras generar la indización, los sistemas de información la incorporan como metadatos para facilitar procesos de descubrimiento, navegación y recomendación. En muchos casos, una sola palabra clave puede representar documentos extensos, lo que convierte a estas unidades en herramientas valiosas para estudios bibliométricos e informétricos que buscan desentrañar la estructura del conocimiento en campos específicos. En estos estudios, la indización, junto con autores o instituciones, permite detectar patrones temáticos y relacionales mediante análisis y visualización de datos.

Varios trabajos ilustran este enfoque. Vargas-Quesada, Chinchilla-Rodríguez y Rodríguez (2017) construyeron un mapa global de la investigación sobre grafeno a partir de más de 50.000 documentos. Su y Lee (2010) realizaron análisis de palabras clave en literatura sobre prospectiva tecnológica, generando visualizaciones en 2D y 3D. Vílchez-Román, Sanguinetti y Mauricio-Salas (2021) usaron términos con alto valor semántico y análisis de co-palabras para apoyar decisiones estratégicas. Cantos Mateos et al. (2013) compararon visualizaciones generadas con tres tipos distintos de descriptores para identificar líneas de investigación. Chávez (2023) desarrolló *MetaMetrics*, un tablero de datos interactivo para evaluar la calidad de los metadatos en revistas científicas. Gálvez (2024) aplicó técnicas de co-palabras sobre palabras clave de autor para analizar la estructura temática de la *Revista General de Información y Documentación*, distinguiendo entre temas centrales, transversales, especializados o emergentes.

Sin embargo, esta visión de la indización enfrenta desafíos en contextos donde el conocimiento cambia rápidamente, como ocurre en los medios informativos. Los vocabularios controlados, concebidos como listas estables y consensuadas, están tensionados por la aparición constante de nuevos temas, entidades y sucesos. Los medios deben responder con rapidez, adaptando sus sistemas de organización del conocimiento de forma ágil y flexible.

Aunque el etiquetado de noticias puede parecer secundario, constituye una parte esencial de la infraestructura mediática. Permite registrar, almacenar y recuperar información mediante estrategias internas como vocabularios propios, bases de datos y metadatos dinámicos. La pandemia de Covid-19 lo ejemplifica: términos como «coronavirus» o «confinamiento», inicialmente útiles, perdieron capacidad discriminatoria al volverse omnipresentes, afectando su valor como términos de indización (Saorín, Pastor-Sánchez y Baños-Moreno 2020).

Los medios, además de seleccionar la noticia, contribuyen a definir qué temas se tornan relevantes. La indización debe reflejar tanto criterios temáticos como su impacto social y temporal. Como advierte Rubio Lacoba (2007), la indización en medios presenta particularidades distintas a otros ámbitos como la ciencia o el entretenimiento.

En este contexto, *El País* ha desarrollado su propio sistema con más de 130.000 términos organizados en categorías como temas, personajes, lugares o eventos (García Jiménez, Rodríguez Mateos y Catalina García 2019). Este vocabulario colaborativo se actualiza con nuevos términos surgidos del discurso social (Rubio Lacoba 2012). Es un sistema híbrido con elementos de tesauro, folksonomía y orientación ontológica, parecido al «organic thesaurus» descrito por Lambe (2007, p. 254). Aunque, como recuerda Broughton (2008), la creación de tesauros es un proceso minucioso que prioriza la estabilidad terminológica, en entornos volátiles como los medios, esta estabilidad puede convertirse en una limitación.

La visualización de información, por su parte, busca identificar y representar gráficamente estructuras significativas dentro de grandes volúmenes de datos. Este proceso implica varias etapas: transformar datos brutos en abstracciones analíticas, construir un modelo visual abstracto y plasmarlo gráficamente mediante principios de diseño (Olmeda-Gómez 2014). Según este autor, la visualización adopta distintas formas: la geoespacial emplea mapas para mostrar información localizada; la jerárquica y de redes representa clasificaciones o relaciones; y la visualización de textos transforma palabras, frases o metadatos en representaciones gráficas.

En suma, la visualización es una herramienta transversal adaptable a cualquier disciplina con datos interpretables, y es cada vez más relevante para la comprensión y comunicación de las realidades sobre las que se aplica (Pérez-Montoro 2025). Su valor reside en convertir información abstracta en representaciones comprensibles, facilitando la detección de patrones y relaciones. Es especialmente útil en contextos con gran volumen de datos, como la ciencia de datos, salud, economía, periodismo, educación, geografía o humanidades digitales, incluyendo la Informetría a través del análisis de palabras clave o descriptores. Actúa como puente entre la sobrecarga informativa y la comprensión intuitiva, facilitando el análisis ágil y significativo.

En este trabajo se utiliza un corpus de indización extenso obtenido de *El País* digital que cubre un periodo de 25 años (2000-2024) para desarrollar análisis estadísticos y métricas que faciliten su exploración visual. El objetivo general es diseñar y validar una aplicación para analizar dinámicamente términos de indización en grandes corpus, con énfasis en su evolución temporal y relacional. Los objetivos específicos son: (1) crear una plataforma interactiva que integre procesamiento estadístico y estructuración semántica; y (2) plantear una metodología para generar métricas que describan el ciclo de vida de los términos, considerando fases como emergencia, estabilización, obsolescencia o recurrencia.

1. METODOLOGÍA

Se ha adoptado un enfoque metodológico de carácter cuantitativo-computacional, sustentado en la extracción sistemática y el análisis de los términos de indización procedentes del archivo de todas las noticias de la edición digital del diario *El País* (2000-2024). La metodología contempla tanto el procesamiento estadístico de los datos como su modelado semántico y su representación visual a través de una aplicación *web* de acceso interactivo. El modelo propuesto se articula en torno a cuatro elementos con un enfoque que permite analizar los términos de indización dentro de un marco replicable en otros corpus similares:

1) Captura y normalización de datos: El proceso comienza con la extracción automatizada de los términos de indización del periódico. Luego, se aplican

técnicas de limpieza, lematización y consolidación de variantes para reducir ambigüedad y garantizar la consistencia de los datos.

2) Modelado y estructuración del *dataset*: Los datos se organizan en un modelo dual: un archivo CSV con métricas estadísticas y un *dataset* RDF. Esto permite un análisis combinado cuantitativo y semántico mediante consultas SPARQL.
3) Elaboración de métricas e indicadores: Se definen indicadores para describir tanto el conjunto como la evolución de cada término: frecuencia, número de fechas, duración de aparición, y medidas derivadas que capturan patrones como irrupciones, estabilidad o decadencia léxica.
4) Exploración visual e interfaz *web*: La información se integra en una aplicación interactiva que ofrece vistas globales y detalladas por término. Incluye filtros temporales, selección y comparación de términos, y visualizaciones dinámicas para facilitar la exploración y el análisis focalizado.

La construcción del conjunto de datos se ha realizado en varias etapas, desde la captura de los datos de noticias y *keywords* del diario *El País*, hasta la construcción del *dataset* final representado mediante RDF. Las etapas han sido las siguientes:

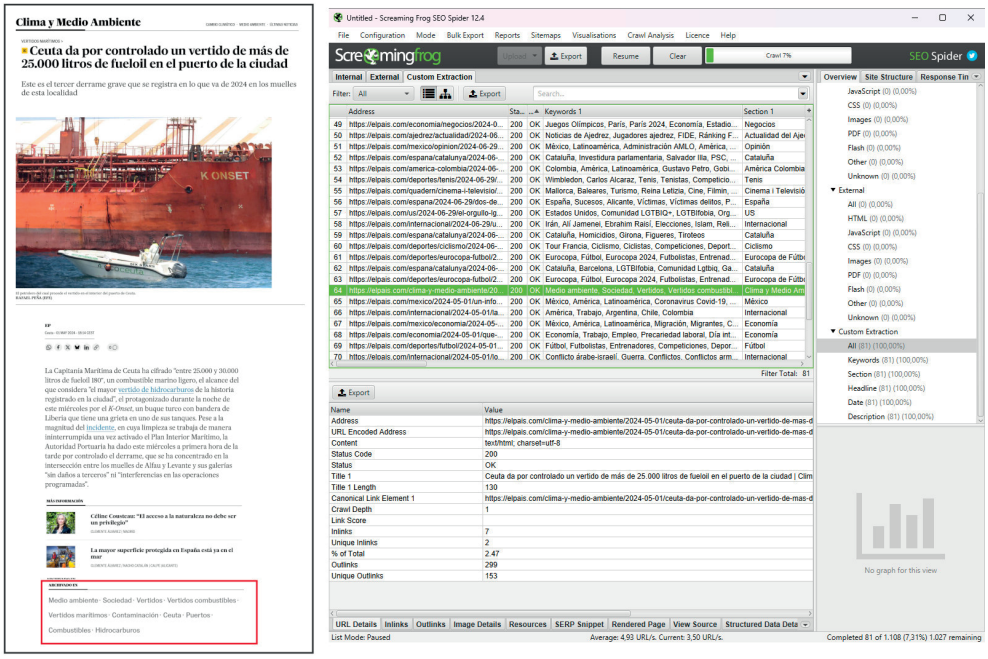


Fig. 2. Extracto de una noticia de *El País* con las palabras clave ubicadas en la parte inferior (marco rojo) y procesamiento de la misma mediante *Screaming Frog SEO*
Fuente: Elaboración propia a partir de Clima… 2024

Etapa 1. Obtención de los datos: Tras realizar un análisis del sitio *web* de *El País* y el marcado de noticias, se aplicaron técnicas de *web scraping* (Alcaraz-Martínez 2025) mediante la aplicación *Screaming Frog SEO*. Como punto de partida se recurrió a los listados diarios de noticias de la hemeroteca[1]. Se realizó un rastreo de todas las fechas anuales conforme a un patrón «http://elpais.com/hemeroteca/<año>-<mes>-<día>/» para el rastreo de URLs de noticias y se excluyeron enlaces a sitios externos o secciones de entretenimiento.

Tras identificar el patrón de URL de las noticias publicadas se excluyeron el resto de páginas de navegación a otras secciones del sitio *web*. De cada noticia se obtuvieron la URL de la página, los términos de indización (bajo el epígrafe «Archivado en»), el titular, la sección y la fecha. Para ello se usaron expresiones regulares y *XPath*. La Tabla 1 muestra un ejemplo de datos extraídos y expresiones regulares utilizadas.

Tabla 1. Ejemplo de datos extraídos y expresiones regulares utilizadas

Elemento	Contenido	Método
Keywords	Economía, Unión Europea, Competencia, Comisión Europea, IAG, Air Europa	Contenido de la etiqueta HTML meta name "news_keywords" Regex: ["']news_keywords["'] content= *\"(.*?)\"
Section	Economía	Contenido de etiqueta HTML meta property "article:section"" Regex: ["']article:section["'] content= *\"(.*?)\"

Fuente: Elaboración propia

Se realizó una exploración separada para cada año usando el patrón de fecha incorporado en las URLs de las noticias, controlando la existencia de duplicados y URLs erróneas. El acceso a la mayoría de las noticias de *El País* se realiza mediante suscripción, pero mediante *web scraping* es posible capturar los metadatos necesarios sin acceder al texto completo. Cada lote supone un tiempo de consulta y procesamiento de entre 6 y 10 horas.

Etapa 2. Unificación y normalización de datos CSV: Mediante *scripts Python*, se realizó la unificación en un único fichero CSV de los múltiples archivos anuales, provenientes de la etapa 1. Sobre este fichero de más de millón y medio noticias se realizó una normalización léxica y tipográfica de los términos de indización recopilados y secciones (transformación de caracteres con tilde, eliminación de caracteres especiales no alfanuméricos y espacios redundantes, conversión a minúsculas, etc.). De este modo, además de los datos de originales (identificador

[1] Puede verse un ejemplo de estas páginas en: *El País…* (2016a; 2016b).

de noticia, URL titular, fecha, secciones y términos de indización) cada sección y término se empareja con su forma normalizada. También se definió para cada noticia un identificador para referenciar de un modo más breve cada noticia en vez de tener que recurrir a la URL original.

Etapa 3. Modelado RDF del conjunto de datos: Consideramos que un aspecto relevante de este trabajo ha sido el modelado RDF del conjunto de datos partiendo del CSV unificado y normalizado. Se utilizaron los vocabularios de *Dublin Core*, *Schema.org* y *Skos*. Para las referencias URL a las noticias, palabras clave y secciones se definieron los correspondientes espacios de nombres. Cada noticia se representar como instancias de la clase *Schema*: *NewsArticle*. Las palabras y conceptos se representaron mediante la clase *Skos*: *Concept*. La Figura 3 muestra el modelo de datos utilizado y la aplicación de las diversas propiedades de los mencionados vocabularios.

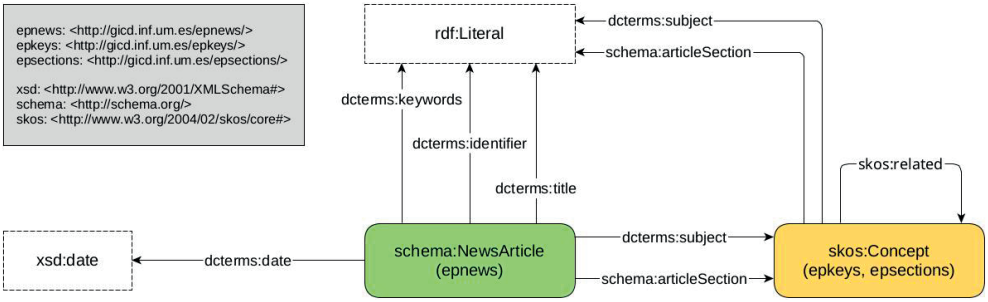


Fig 3. Diagrama del modelo de datos de representación y espacios de nombres utilizados
Fuente: Elaboración propia

Etapa 4. Cálculo de estadísticas básicas de palabras clave, métricas agregadas y *clustering*: Tras unificar el *dataset* en un fichero CSV se calcularon estadísticas del uso de los términos de indización. Se identificaron también aquellos términos que tienen la misma denominación que las secciones. Para cada término representado se calcularon tres dimensiones: número total de noticias que indiza, número de días únicos en los que se usa y el número de días del intervalo de uso temporal. Estas tres dimensiones se transformaron logarítmicamente y se utilizaron con dos propósitos. El primero fue la representación de cada término como un vector cuyas componentes son las mencionadas componentes. El segundo propósito fue la obtención de un *clustering* mediante la aplicación del algoritmo *K-Means* para obtener una clasificación de los términos en función de su uso.

2. RESULTADOS

Tras procesar los datos de 1.605.589 noticias de prensa de 25 años, que abarcan desde el 1 de enero de 2000 hasta el 31 de diciembre de 2024, se obtuvo un conjunto de datos RDF con 41.997.261 tripletas. Las noticias se organizan en 677 secciones. El total de palabras clave usadas es de 73.392 que se han utilizado en 17.275.146 ocasiones para la indización de noticias. Las palabras clave se tratan como entidades independientes, tal y como se publican en la *web*, sin relaciones jerárquicas, ni variantes de términos preferentes ni otras relaciones habituales en vocabularios controlados para la indización temática.

Por su parte, los resultados del proceso de *clustering* arrojan unos resultados óptimos para cinco grupos de palabras clave. La Figura 4 ofrece una representación y su distribución tridimensional según el número total de noticias que indiza, el número de días únicos en los que se usa y el número de días del intervalo de uso temporal.

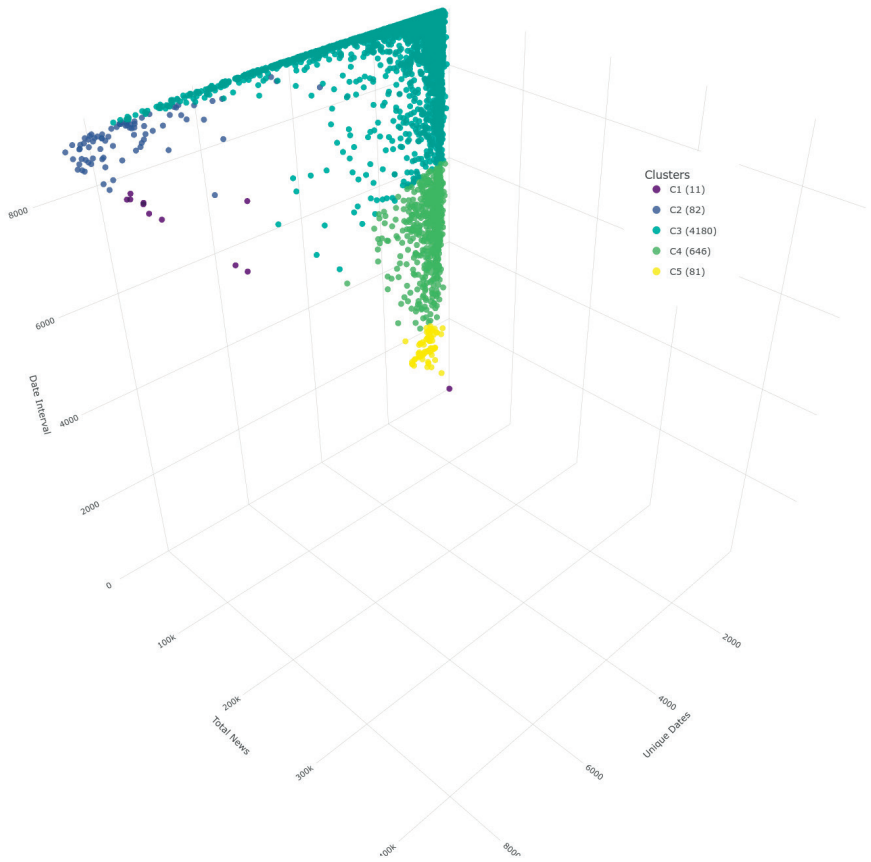


Fig. 4. Resultados del *clustering* mediante el algoritmo *K-Means*
Fuente: Elaboración propia

En el *clúster* C1 pueden encontrarse términos como «España», «Economía», «Cultura», «Sociedad», etc.: se trata de palabras clave que más bien se corresponden con secciones del periódico. En C2 hay términos muy generales que se aplican en gran cantidad de noticias como «arte», «gobierno» o «problemas sociales» entre otros. C3 agrupa términos de un uso relativamente frecuente como «calor», «infracciones urbanísticas», «desempleo» o «desastres naturales». C4 incluye términos muy vinculados con nombres de personas, lugares o eventos concretos. Por su parte el *clúster* C5 alberga palabras clave marginales, a veces de uso único, y que hacen frente a necesidades muy puntuales y con poco recorrido temporal.

Se ha desarrollado un prototipo de aplicación *web*[2], que permite acceder al análisis del *corpus* de términos de indización usados en las noticias. Se persigue ofrecer una interfaz que permita visualizar la presencia y evolución temporal de términos mediante gráficos, tablas y representaciones tridimensionales. Además del tratamiento de datos tabulados, la aplicación se conecta a un conjunto de datos RDF accesible mediante consultas SPARQL, lo que permite integrar información semántica adicional sobre cada concepto.

El sistema utiliza tecnologías de desarrollo *web* de propósito general (PHP, JavaScript, HTML, CSS), junto con bibliotecas especializadas para visualización (Plotly.js, DataTables). Su diseño modular permite que los usuarios naveguen entre diferentes tipos de vistas: desde estadísticas generales y listados completos hasta análisis detallados por término o exploración semántica mediante grafos de coocurrencia. y permite:

- La exploración interactiva de términos asociados a noticias del diario *El País*.
- El análisis temporal, cuantitativo y contextual de estos conceptos durante 25 años.
- La representación visual de métricas como frecuencia, dispersión o recurrencia.
- La navegación entre vistas globales y detalladas, incluyendo la coocurrencia.
- La integración de datos tabulados (CSV) y estructurados (RDF) en un mismo entorno.

Los datos provienen de dos fuentes principales:

1) Archivo CSV: contiene la forma canónica de cada palabra clave, la frecuencia de aparición total, el número de fechas distintas en que aparece y la duración del intervalo en días entre la primera y última mención. Este archivo es procesado por funciones PHP que lo transforman en estructuras asociativas para su posterior filtrado o visualización.

[2] El prototipo se encuentra disponible en: *Dataset El País (2000-2024)* [En línea] [consult. 2025-05-10]. Disponible en: https://gicd.inf.um.es/kw.

2) Grafo RDF: consultado mediante SPARQL, proporciona información semántica sobre los conceptos: etiquetas, propiedades, relaciones jerárquicas y equivalencias terminológicas. Estas consultas se realizan en tiempo real y los resultados se integran dinámicamente en las vistas.

Los datos del CSV se procesan en el servidor para aplicar filtros por fecha, generar subconjuntos o calcular métricas derivadas (por ejemplo, logaritmos o *rankings*). El sistema permite conservar la estructura modular de los datos, actualizándolos fácilmente sin modificar el código.

El sistema se basa en una arquitectura cliente-servidor. En el servidor, los datos se cargan desde un archivo CSV que contiene información estadística por concepto: frecuencia total, fechas de aparición y duración del uso. Estos datos son procesados con funciones PHP y enviados al cliente en formato JSON. Desde el lado cliente, la visualización se realiza mediante HTML y JavaScript, utilizando Plotly.js para los gráficos y DataTables para las tablas. Además, se integra un *endpoint* SPARQL que permite internamente a la aplicación consultar un grafo RDF asociado a los conceptos, recuperando información semántica adicional como etiquetas alternativas, relaciones jerárquicas o agrupaciones temáticas. Esta dualidad permite combinar análisis cuantitativo y estructural.

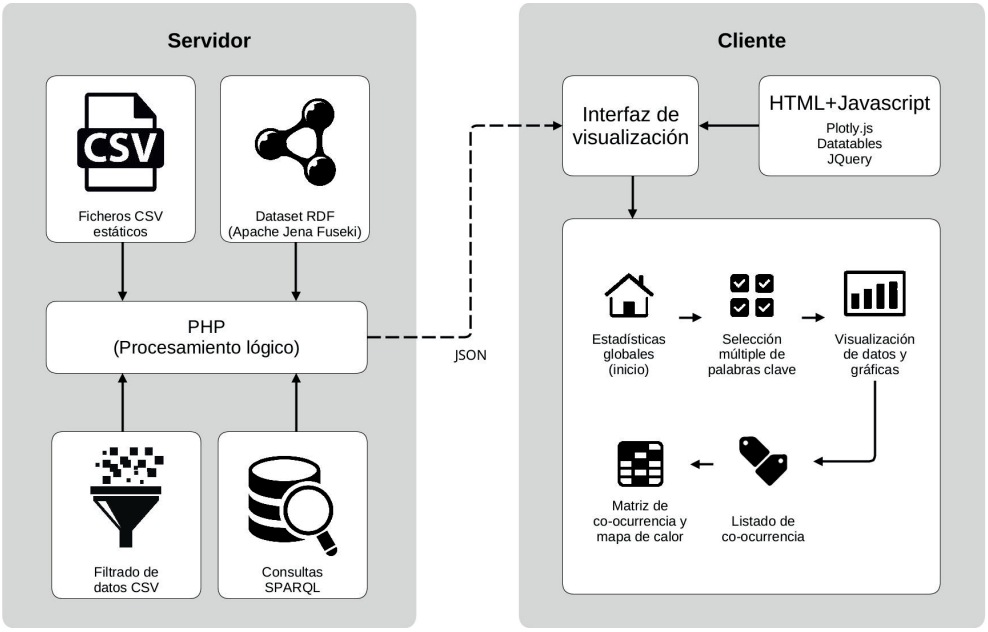


Fig. 5. Arquitectura y flujo de uso de la aplicación
Fuente: Elaboración propia

El usuario accede a la página de inicio, donde se presentan estadísticas generales del corpus de indización: número total de noticias, volumen de términos de indización y distribución temporal. Un histograma muestra la cantidad de noticias por año, mientras que un gráfico de dispersión 3D permite explorar la distribución estadística de conceptos según métricas como frecuencia, dispersión temporal y número de fechas únicas.

Desde allí, puede dirigirse a la sección «KEYWORDS» del menú, donde se presenta una tabla con el listado completo de conceptos. Esta tabla incluye funciones de búsqueda, ordenación y selección múltiple. El usuario puede filtrar conceptos por término o raíz, y seleccionar un conjunto concreto para análisis comparativo.

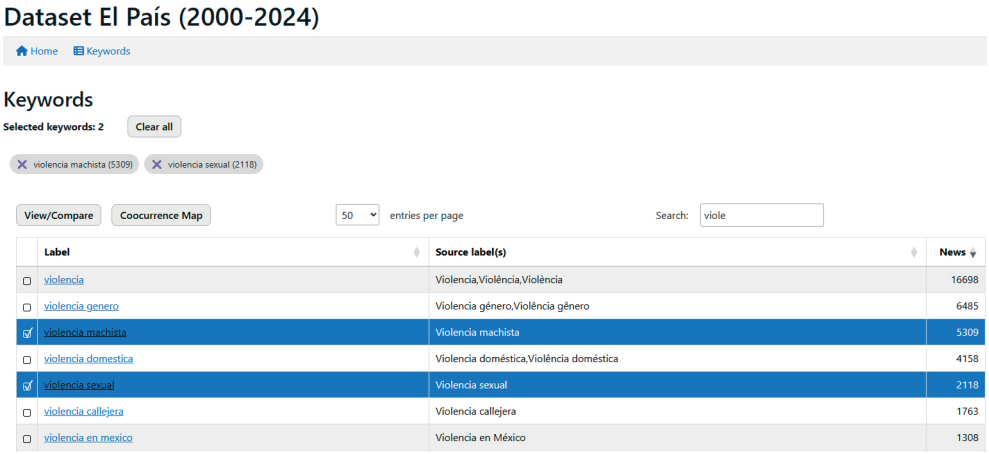


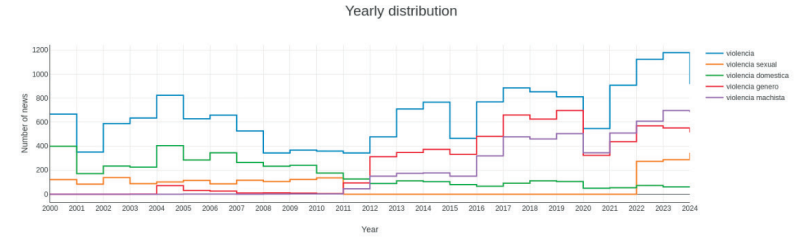
Fig. 6. Filtrado y selección de términos en la aplicación
Fuente: Elaboración propia

Una vez seleccionados los términos, el usuario puede elegir entre las opciones «View/Compare» y «Coocurrence Map» para el conjunto de términos escogidos. Desde «View/Compare» accede a un panel de visualización temporal. Este módulo presenta distintos gráficos comparativos: distribución anual, frecuencias absolutas por intervalo, diferencias de frecuencia y tablas de correlación. También permite analizar periodos de mayor presencia y explorar relaciones estadísticas entre conceptos.

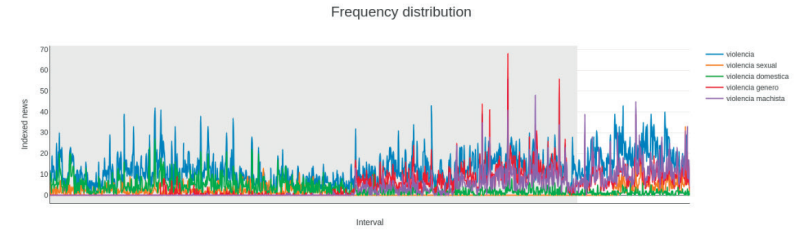
Dataset El País (2000-2024)

Home | Keywords

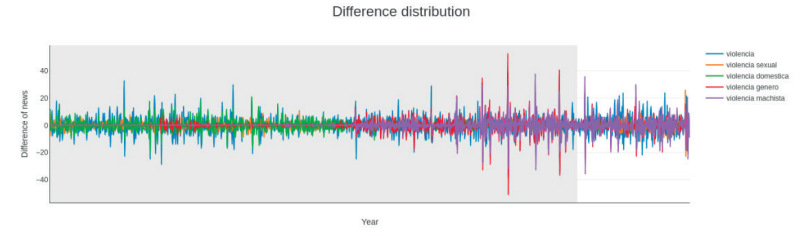
Keyword label	Keyword sources	Main detected period	Indexed news	Co-Keywords
violencia	Violencia, Violência, Violència	130.33	16698	7675
violencia domestica	Violencia doméstica, Violência doméstica	-	4158	2831
violencia genero	Violencia género, Violência gênero	-	6485	4559
violencia machista	Violencia machista	-	5309	3858
violencia sexual	Violencia sexual	45.5	2118	2131



Normalized values



Normalized values



Absolute values | Normalized values | IDTF values | IDTF normalized

Correlations

	violencia	violencia sexual	violencia domestica	violencia genero	violencia machista
violencia	1	0.16083160827013	0.26418148646142	0.56316126305462	0.53372413341902
violencia sexual	0.16083160827013	1	0.067948783998307	0.035951749059554	0.23441353754777
violencia domestica	0.26418148646142	0.067948783998307	1	-0.24021306230661	-0.28060058891644
violencia genero	0.56316126305462	0.035951749059554	-0.24021306230661	1	0.88971088468301
violencia machista	0.53372413341902	0.23441353754777	-0.28060058891644	0.88971088468301	1

Cross-Correlations

	violencia	violencia sexual	violencia domestica	violencia genero	violencia machista
violencia	1 (lag: 0)	0.177256 (lag: 9)	0.264181 (lag: 0)	0.563161 (lag: 0)	0.533724 (lag: 0)
violencia sexual	0.177256 (lag: -9)	1 (lag: 0)	0.076203 (lag: 1)	-0.080999 (lag: 9)	0.234414 (lag: 0)
violencia domestica	0.264181 (lag: 0)	0.076203 (lag: -1)	1 (lag: 0)	-0.336718 (lag: -10)	-0.333781 (lag: 7)
violencia genero	0.563161 (lag: 0)	-0.080999 (lag: -9)	-0.336718 (lag: 10)	1 (lag: 0)	0.889711 (lag: 0)
violencia machista	0.533724 (lag: 0)	0.234414 (lag: 0)	-0.333781 (lag: -7)	0.889711 (lag: 0)	1 (lag: 0)

Fig. 7. Datos comparados de distribución de frecuencias de uso de palabras clave y datos de correlación entre las palabras clave
Fuente: Elaboración propia

En el mismo módulo, el usuario puede seleccionar un término individual y acceder a su vista detallada. Allí se presentan metadatos (como etiquetas fuente, fechas de aparición y número total de noticias), así como un listado de *co-keywords* con los que aparece con más frecuencia.

Desde esta misma vista se ofrece la opción de consultar el mapa de coocurrencias. El usuario elige un término principal y se genera automáticamente una tabla de coapariciones por año con otros conceptos seleccionados, junto con un mapa de calor que representa gráficamente la intensidad de dichas coocurrencias a lo largo del tiempo.

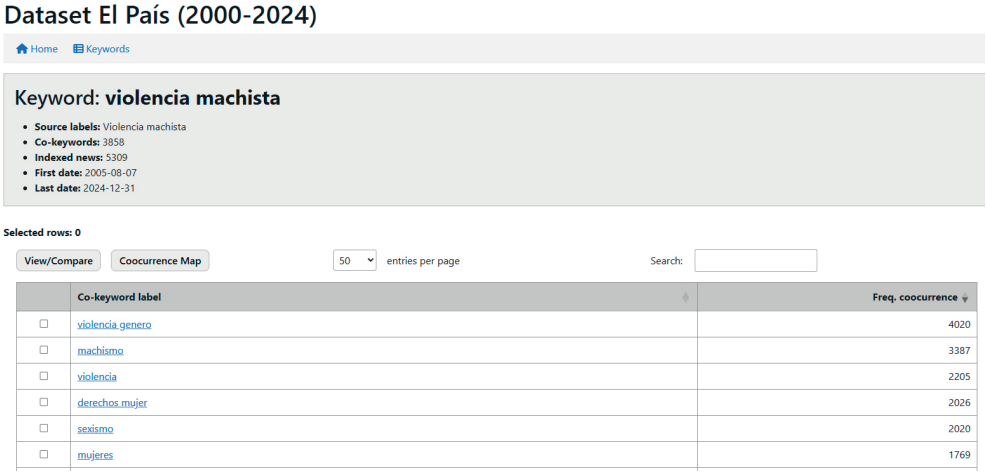


Fig. 8. Metadatos y coocurrencias de un término
Fuente: Elaboración propia

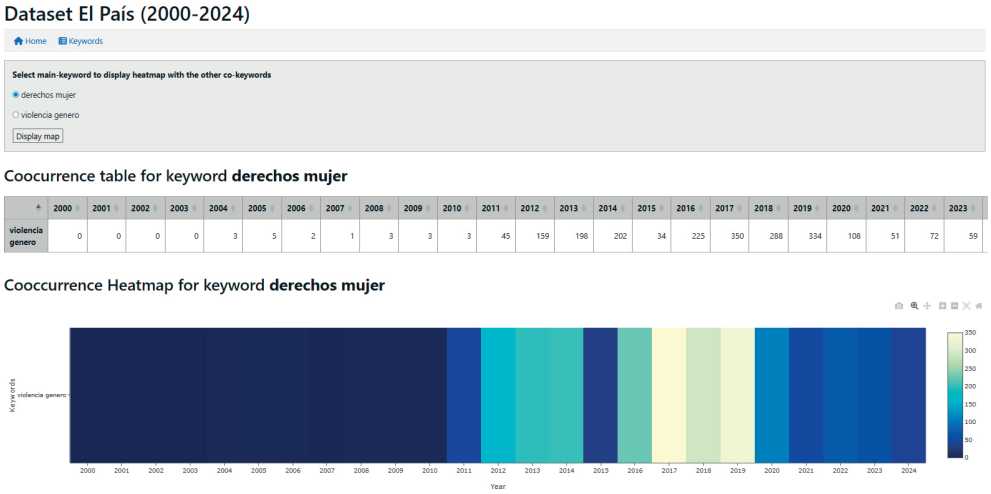


Fig. 9. Mapa de calor de coocurrencias de un término principal
Fuente: Elaboración propia

3. DISCUSIÓN: HACIA UN MODELO DE MÉTRICAS E INDICADORES DE INDIZACIÓN

Además de la creación del corpus descrito se propone una primera aproximación a un modelo de métricas e indicadores para comprender mejor la elaboración de vocabularios y la práctica de la indización, mediante su explotación estadística y visual.

Las infraestructuras de análisis de la producción científica tienen un propósito vinculado a la evaluación de la ciencia, la jerarquización de sus resultados y el análisis de redes de citas y colaboración. Permiten el mapeado temático y la detección de temas emergentes. Dado que muchas herramientas de análisis temático se aplican sobre texto completo, su uso en la indización controlada exige ciertas especificidades.

En este marco, se definen métricas e indicadores que permiten examinar cuantitativamente el uso y evolución del vocabulario en el corpus. La propuesta distingue dos niveles: general del corpus e individual para cada término. Se parte de la diferencia entre métricas (valores observables) e indicadores (combinaciones interpretativas), articulados en torno a dimensiones analíticas. Una misma métrica puede participar en más de una dimensión:

1) Frecuencia de uso: Cuenta cuántas veces aparece un término. Distingue términos comunes y marginales, y sirve como base para otros análisis.
2) Temporalidad y actividad: Analiza cuándo y cuánto tiempo se usa un término. Identifica persistencia, estacionalidad o irrupciones.
3) Densidad de uso: Observa la intensidad del uso a lo largo del tiempo. Permite detectar fenómenos como explosividad o dispersión.
4) Conectividad semántica: Evalúa la coocurrencia con otros términos. Una alta conectividad sugiere redes conceptuales amplias.
5) Rareza y especificidad: Detecta términos muy poco frecuentes, que pueden ser especializados o errores.
6) Evolución léxica: Estudia los cambios terminológicos, como la aparición de nuevos términos, relevante en contextos dinámicos.
7) Cobertura documental: Examina cómo se distribuyen los términos entre documentos, revelando posibles desequilibrios.

A escala global, estas dimensiones permiten abordar aspectos dinámicos del vocabulario y su ciclo de vida. Métricas como el número medio y la mediana de términos por documento reflejan la densidad de anotación y posibles sesgos. La extensión del vocabulario y la frecuencia media por término indican riqueza terminológica y nivel de reutilización.

También se mide el total de apariciones y cuántas son únicas. Estos 'términos raros' pueden reflejar especificidad o inconsistencias. Se derivan indicadores como el índice de contenido exclusivo o el índice de reutilización, que evalúan cómo se comparte el vocabulario y su eficiencia.

La dimensión temporal es clave. El análisis del número medio de días activos por término, o la duración entre primera y última aparición, permite calcular el grado de concentración temporal. Esto revela si un término aparece de forma sostenida o en un corto lapso. También se mide la dispersión temporal y la estabilidad, así como la media de nuevos términos por día, útil para observar el crecimiento y transformación del vocabulario.

En el plano semántico, se analizan coocurrencias. Saber qué términos aparecen juntos permite identificar asociaciones conceptuales. Se calculan métricas como el número medio de términos con los que coocurre cada uno, y se derivan indicadores como conectividad y densidad semántica, que revelan centralidad conceptual.

También se estudia la intensidad del uso. Algunos términos aparecen esporádicamente, otros de forma muy concentrada. Métricas como la frecuencia media por día activo o los días inactivos entre apariciones permiten calcular el índice de explosividad, que mide si el uso es sostenido o abrupto.

A nivel de palabra clave individual, se calculan valores como frecuencia documental, duración de uso o rareza inversa. Este último destaca términos extremadamente infrecuentes. También se incluye el índice de significancia, que combina frecuencia, foco temporal y conectividad para estimar el valor analítico de un término.

La evolución del repertorio terminológico se analiza con la métrica de nuevos términos por día, que permite detectar especialización y transformaciones discursivas en corpus en expansión.

Las Tablas 2 y 3 recogen un inventario de las métricas e indicadores propuestos, su codificación, cálculo. Se diferencia entre las que son obtenidas a nivel de global de conjunto de datos y las de nivel de términos (palabras clave). Al final de cada tabla se incluye un listado descriptivo de cada elemento del modelo.

Tabla 2. Métricas e indicadores a nivel global del *dataset*

Código	Nombre	Cálculo
GM1	total_docs	COUNT(doc)
GM2	keyword_vocab_size	COUNT(keyword)
GM3	total_keyword_occurrences	SUM(KM1)
GM4	avg_keywords_per_doc	GM3 / GM1
GM5	avg_doc_freq	SUM(KM1) / GM2

(*continua na página seguinte*)

Código	Nombre	Cálculo
GM6	avg_active_days	SUM(KM5) / GM2
GM7	avg_timespan_days	SUM(KM4) / GM2
GM8	rare_keywords	COUNT(keyword:doc_freq=1)
GD1	long_tail_ratio	GM8 / GM2
GD2	keyword_reuse_ratio	GM3 / GM2
GM9	avg_cooc_keywords	SUM(KM10) / GM2
GM10	keywords_per_day_avg	SUM(KM1) / SUM(KM5)
GM11	docs_per_day_avg	GM1 / GM7

Fuente: Elaboración propia

- GM1: Total de documentos del *corpus*. Sirve como base para contextualizar las demás métricas y evaluar la intensidad de indización.
- GM2: Número de *keywords* distintas. Mide la riqueza y diversidad semántica del vocabulario usado.
- GM3: Total de asignaciones de términos. Indica la carga global de anotación: valores altos reflejan alta densidad; valores bajos reflejan escasa indización.
- GM4: Media de términos por documento. Evalúa la densidad semántica: valores altos implican detalle; valores bajos indican un etiquetado superficial.
- GM5: Frecuencia media de aparición de cada término en documentos. Mide su reutilización: valores altos indican un vocabulario central; valores bajos indican dispersión terminológica.
- GM6: Media de días en que los términos están activos. Informa sobre la estabilidad temporal: media alta apunta a un léxico persistente; media baja se refiere posiblemente a una aparición efímera.
- GM7: Duración media entre primera y última aparición de los términos. Indica su vida útil: alta, uso sostenido; baja, episodios puntuales.
- GM8: Número de términos que aparecen solo una vez. Refleja rareza: valores altos indican vocabulario específico o poco estandarizado.
- GM9: Media de coocurrencias por término. Mide conectividad semántica: valores altos, vocabulario relacional; bajos, términos aislados.
- GM10: Términos activos por día. Indica vitalidad semántica diaria: alta, riqueza y dinamismo; baja, renovación léxica escasa.
- GM11: Documentos generados por día. Mide la dinámica de crecimiento: alta, productividad sostenida; baja, ritmo lento.
- GD1: Indica cuántas *keywords* se usan solo una vez. Un valor alto sugiere diversidad o falta de control; un valor bajo, vocabulario más reutilizado.
- GD2: Mide la reutilización media del vocabulario. Un valor alto refleja eficiencia; uno bajo, dispersión o escasa coherencia.

Tabla 3. Métricas e indicadores a nivel global de palabras clave

Código	Nombre	Cálculo
KM1	doc_freq	COUNT(doc:keyword)
KM2	first_date	MIN(date:doc:keyword)
KM3	last_date	MAX(date:doc:keyword)
KM4	timespan_days	KM3-KM2
KM5	active_days	COUNT(DISTINCT(date:doc:keyword)
KM6	avg_freq_per_day	KM1/KM4
KM7	usage_density	KM1/KM5
KM8	gap_days_avg	(KM4-KM5)/(KM5-1)
KM9	burstiness_index	KM7/KM6
KM10	cooc_keywords	COUNT(DISTINCT(keyword:keyword)
KD1	temporal_focus	KM5/KM4
KD2	temporal_sparsity	1-KD1
KD3	keyword_stability	KM5/KM1
KD4	semantic_connectivity	KM10/KM1
KD5	semantic_density	KM10/KM5
KD6	rarity_index	1/KM1 o alternativamente LOG(1+1/KM1)
KD7	burst_score	KM9 * (1-KD1)
KD8	significance_index	$SQRT(LOG(1+KM1)^2 + KD1^2 + LOG(1+KM10)^2)$

Fuente: Elaboración propia

- KM1: Número de documentos en que aparece un término. Un valor alto indica uso generalizado; uno bajo, especialización o marginalidad.
- KM2: Fecha de primera aparición. Una fecha temprana sugiere vocabulario fundacional; una reciente, innovación o incorporación tardía.
- KM3: Fecha de última aparición. Si es reciente, el término sigue vigente; si es antigua, puede estar en desuso.
- KM4: Días entre primera y última aparición. Un valor alto indica larga trayectoria; uno bajo, uso puntual o efímero.
- KM5: Número de días únicos con uso del término. Un valor alto muestra constancia; uno bajo, uso ocasional.
- KM6: Frecuencia media diaria en su periodo de uso. Alto valor implica intensidad sostenida; bajo, relevancia limitada.
- KM7: Frecuencia media en días activos. Un valor alto indica concentración puntual; uno bajo, distribución más uniforme.
- KM8: Media de días inactivos entre usos. Alto valor sugiere intermitencia; bajo, regularidad.

- KM9: Índice de explosividad. Un valor alto señala irrupción intensa; uno bajo, uso más estable.
- KM10: Número de términos con los que coocurre. Un valor alto refleja conectividad semántica; uno bajo, aislamiento.
- KD1: Proporción de días activos en su vida útil. Alto valor implica uso concentrado; bajo, disperso.
- KD2: Dispersión temporal (inverso de KD1). Valor alto indica uso irregular; bajo, continuidad.
- KD3: Relación entre días activos y frecuencia. Un valor alto indica estabilidad; uno bajo, uso agrupado.
- KD4: Coocurrencias por aparición. Alto valor denota riqueza contextual; bajo, uso limitado o aislado.
- KD5: Coocurrencias por día activo. Alto valor indica riqueza conceptual; bajo, posible especialización extrema.
- KD6: Rareza inversa del término. Un valor alto lo identifica como poco frecuente; bajo, como común.
- KD7: Explosividad ponderada por dispersión. Alto valor refleja irrupciones temáticas; bajo, evolución gradual.
- KD8: Índice compuesto de frecuencia, foco y conectividad. Un valor alto indica importancia global del término; bajo, escasa relevancia.

La Tabla 4 muestra a qué dimensión del modelo se vincula cada una de las métricas e indicadores definidos, ya sea a nivel global o a nivel de palabra clave.

Tabla 4. Dimensiones del Modelo y sus métricas e indicadores asociados

Dimensión	Métricas asociadas	Indicadores asociados
Frecuencia de uso	GM1, GM2, GM3, GM4, GM5, KM1	GD2
Temporalidad y actividad	GM6, GM7, GM10, GM11, KM2, KM3, KM4, KM5	KD1, KD2, KD3, KD7, KD8
Densidad de uso	KM6, KM7, KM8	KM9
Conectividad semántica	GM9, KM10	KD4, KD5, KD8
Rareza y especificidad	GM8	GD1, KD6
Evolución léxica	GM12	-
Cobertura documental	GM4, GM11	-

Fuente: Elaboración propia

CONCLUSIONES

Este trabajo presenta un enfoque híbrido que combina la construcción conceptual de un modelo con la validación empírica a través de un caso real. La propuesta metodológica se muestra robusta y transferible a otros contextos (colecciones institucionales, científicas o educativas). Se está avanzando en el diseño de una herramienta accesible públicamente que permita importar nuevos *datasets* de indización y obtener automáticamente indicadores y visualizaciones, facilitando así la comprensión de la evolución terminológica en diferentes ámbitos de uso y dominios del conocimiento. Además, esta herramienta pretende fomentar el análisis comparado entre vocabularios y su evolución en el tiempo, fortaleciendo la investigación sobre la organización del conocimiento desde una perspectiva empírica y exploratoria; y que pudiera ser de utilidad para propiciar investigaciones en otras áreas de conocimiento.

La propuesta que presentamos aquí se sitúa en la intersección entre los estudios sobre vocabularios controlados y el análisis cuantitativo de su uso en contextos reales, apoyado en técnicas de visualización de la información. Partiendo de un gran corpus de noticias del diario *El País* (España) que cubre un periodo de 25 años, se propone un modelo para estudiar y visualizar la evolución de los términos de indización, con especial atención a su ciclo de vida, patrones de uso y relaciones semánticas. La propuesta parte de una concepción dinámica de los vocabularios, como entidades vivas que responden a fenómenos sociales, políticos y culturales. Este enfoque permite observar cómo los términos se adaptan, emergen o desaparecen en función del contexto informativo, reflejando así los cambios en la agenda pública y en los marcos de referencia sociales.

Una modelización de estas características ayudaría a hacer más visible lo invisible y podría tener varias implicaciones. Por un lado implicaciones para los propios medios de comunicación, puesto podrían realizar análisis de tendencias temáticas a lo largo del tiempo y comprobar los cambios en los intereses informativos (como por ejemplo, observar cómo la expresión «cambio climático» dejó de ser un simple concepto para ser una sección propia); detección de sesgos editoriales o zonas de silencio (si ciertos temas están ausentes o sobrerrepresentados); o estrategias de contenido y posicionamiento (identificación de nichos temáticos con baja cobertura pero alta relevancia emergente).

Paralelamente, lo que ha sido mostrado presenta perspectivas aplicables en diferentes disciplinas académicas. En Comunicación y Periodismo, pueden enriquecer los estudios de agenda *setting* (los temas importantes para los medios) y *framing* (desde la perspectiva que se cuentan); análisis de cobertura mediática y representación social, evolución del tratamiento de ciertos temas, estudio de transformación del lenguaje, etc. En Sociología, podría aplicarse en los análisis de cambios socioculturales reflejados en los medios, o el estudio de fenómenos como movimientos sociales,

identidad, globalización, etc. Igualmente, en los estudios culturales y de antropología, para la observación de cómo se construyen narrativas culturales en los medios. Y por último, para el área de Biblioteconomía y Documentación y en especial en lo relativo a tareas de mantenimiento y actualización de vocabularios controlados, una modelización de este tipo podría resultar de interés para conocer la evolución terminológica mediante la identificación de conceptos emergentes, terminología obsoleta u observar sinonimias o variantes léxicas; así como para mejorar el control terminológico, analizando la coherencia en la asignación de los términos o tal vez la consistencia terminológica del sistema de indización. Máximo teniendo en cuenta que la indización en el periódico *El País* la realizan tanto los mismos periodistas como documentalistas. Por último, en el ámbito docente, esta herramienta también podría emplearse para actividades prácticas relacionadas con análisis textual, visualización de datos, estudio de coocurrencias o trabajo con ontologías. El hecho de basarse en tecnologías *web* comunes facilita su integración en entornos educativos y su uso sin requerimientos técnicos complejos.

Entre las líneas de desarrollo previstas para este proyecto se contempla la incorporación de nuevos tipos de visualizaciones (como redes de coocurrencias o grafos semánticos), delimitar análisis por periodos temporales establecidos por el usuario, la mejora de la exportación de resultados, la ampliación del corpus a otros medios o idiomas y la optimización del sistema para uso multiusuario. También se prevé ampliar la integración con fuentes de datos externas o sistemas de metadatos interoperables. En especial se contempla enriquecer el *dataset* mediante la tipificación de los términos de indización desde el punto de vista del Reconocimiento de Entidades Nombradas (NER), para estudiar por separado personas, organizaciones, lugares o temas. El sistema se mantiene como una herramienta adaptable, orientada a la exploración de dinámicas conceptuales en corpus extensos de indización y categorización, y puede aplicarse en contextos académicos, institucionales o educativos según las necesidades de análisis y visualización. Otra línea en la que trabajaremos será en la creación de estudios de caso en los que a partir de un acontecimiento o término se puedan explicar las métricas, indicadores y aspectos relevantes para comprender mejor la indización desde la *praxis* y la semántica.

Vivimos una explosión del interés en los estudios analíticos sobre el lenguaje, con un sinfín de aproximaciones metodológicas y librerías de *software*. Es relevante que también nos aproximemos a los lenguajes controlados con nuevas herramientas no solo para observar los resultados, sino también para afinar la *praxis* de la indización (humana y automatizada) y dar soporte a los procesos de construcción de vocabularios y organización de taxonomías.

REFERENCIAS

ALCARAZ-MARTÍNEZ, Rubén, 2025. Beautiful Soupde Python para el raspado web como método para la extracción automatizada de datos. *Infonomy* [En línea]. **2**(3), e25014. ISSN: 2990-2290 [consult. 2025-05-10]. Disponible en: https://infonomy.scimagoepi.com/index.php/infonomy/article/view/95.

BROUGHTON, John, 2008. *Wikipedia: The Missing Manual*. [S.l.]: O'Reilly Media. ISBN 978-0596515164.

CANTOS MATEOS, Gisela, et al., 2013. Estudio comparativo sobre la visualización de redes de co-words a través de los descriptores del Science Citation Index y de Medline. En: *Informação e/ou Conhecimento: as duas faces de Jano: I Congresso ISKO Espanha e Portugal / XI Congreso ISKO España. Atas.* Porto: Faculdade de Letras da Universidade do Porto, ISKO, pp. 173-189.

CHÁVEZ, Manuel, 2023. MetaMetrics: prototipo de visualización de la calidad de los metadatos en revistas científicas latinoamericanas publicadas en Open Journal System. *Biblioteca Universitaria: Revista de la Dirección de Bibliotecas de la UNAM* [En línea]. **26**(1), 12-23 [consult. 2025-05-10]. ISSN 0187-750X. Disponible en: https://bibliotecauniversitaria.dgb.unam.mx/rbu/article/view/1466/1425.

Clima y Medio Ambiente. Ceuta da por controlado un vertido de más de 25.000 litros de fueloil en el puerto de la ciudad. *El País* [En línea]. 2024-05-01 [consult. 2025-05-10]. Disponible en: https://elpais.com/clima-y-medio-ambiente/2024-05-01/ceuta-da-por-controlado-un-vertido-de-mas-de-25000-litros-de-fueloil-en-el-puerto-de-la-ciudad.html.

El País. Hemeroteca [En línea]. 2016a. España: El País [consult. 2025-05-10]. Disponible en: https://elpais.com/hemeroteca/2016-01-30/.

El País. Hemeroteca [En línea]. 2016b. España: El País [consult. 2025-05-10]. Disponible en: https://elpais.com/hemeroteca/2016-01-30/2/.

GÁLVEZ, C., 2024. Mapas científicos de la Revista General de Información y Documentación (2005-2022). *Revista General de Información y Documentación* [En línea]. **34**(1), 127-140 [consult. 2025-05-10]. Disponible en: https://revistas.ucm.es/index.php/RGID/article/view/88515.

GARCÍA JIMÉNEZ, Antonio, David RODRÍGUEZ MATEOS, y Beatríz CATALINA GARCÍA, 2019. Estudio sobre la indización/etiquetado y los lenguajes documentales en cinco diarios españoles. *Scire: Representación Y organización Del Conocimiento* [En línea]. **25**(1), 55-64 [consult. 2025-05-10]. Disponible en: https://www.ibersid.eu/ojs/index.php/scire/article/view/4579.

GIL-LEIVA, Isidoro, 2025. Indexing. En: David BAKER, y Lucy ELLIS, ed. *Encyclopedia of Libraries, Librarianship, and Information Science* [En línea]. [S.l.]: Academic Press, vol. 2, pp. 514-531 [consult. 2025-05-10]. [S.l.]: Elsevier. DOI: http://dx.doi.org/10.1016/B978-0-323-95689-5.00205-4.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 1985. *ISO 5963-1985: Documentation – Methods for Examining Documents, Determining Their Subjects, and Selecting Indexing Terms*. Geneva: ISO.

LAMBE, Patrick, 2007. *Organising knowledge: taxonomies, knowledge and organizational effectiveness*. Oxford: Chandos Publishing. ISBN 9781843342274.

OLMEDA-GÓMEZ, Carlos, 2014. Visualización de información. *El profesional de la información* [En línea]. **23**(3), 213-219 [consult. 2025-05-10]. Disponible en: https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/epi.2014.may.01.

PÉREZ-MONTORO, M., 2025. Annual Overview of the Field of Information Visualization. *Hipertext.net* [En línea]. 30, 203-226 [consult. 2025-05-10]. DOI: https://doi.org/10.31009/hipertext.net.2025.i30.13.

RUBIO LACOBA, María, 2012. Nuevas destrezas documenta-les para periodistas: el vocabulario colaborativo del diario El País. *Trípodos*. 31, 65-78.

RUBIO LACOBA, María, 2007. *Documentación informativa en el periodismo digital.* Madrid: Síntesis.

SAORÍN, Tomás, Juan-Antonio PASTOR-SÁNCHEZ, y María José BAÑOS-MORENO, 2020. Uso de Wikidata y Wikipedia para la generación asistida de un vocabulario estructurado multilingüe sobre la pandemia de Covid-19. *Profesional de la información* [En línea]. **29**(5), e290509 [consult. 2025-05-10]. DOI: https://doi.org/10.3145/epi.2020.sep.09.

SU, Hsin-Ning y Pei-Chun LEE, 2010. Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight. *Scientometrics*. **85**(1), 65-79.

VARGAS-QUESADA, Benjamín, Zaida CHINCHILLA-RODRÍGUEZ, y Noel RODRIGUEZ, 2017. Identification and Visualization of the Intellectual Structure in Graphene Research. *Frontiers in Research Metrics and Analytics* [En línea]. **2**(oct.), Article 7 [consult. 2025-05-10]. DOI: https://doi.org/10.3389/frma.2017.00007.

VÍLCHEZ-ROMÁN, Carlos, Sol SANGUINETTI, y Mariela MAURICIO-SALAS, 2021. Applied bibliometrics and information visualization for decision-making processes in higher education institutions. *Library Hi Tech*. **39**(1), 263-283.