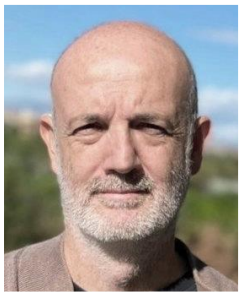
# Las palabras clave se agitan con el viento: explorando la dinámica de uso de descriptores temáticos en El País para las noticias sobre “amnistía” durante 25 años

## Keywords move with the wind: exploring the dynamics of use of thematic descriptors in El País for news about “amnesty” over 25 years

**Tomás Saorín-Pérez; Juan-Antonio Pastor-Sánchez**

Cómo citar este artículo:

**Tomás, Saorín-Pérez; Pastor-Sánchez, Juan-Antonio** (2024). “Las palabras clave se agitan con el viento: explorando la dinámica de uso de descriptores temáticos en El País para las noticias sobre “amnistía” durante 25 años [Keywords move with the wind: exploring the dynamics of use of thematic descriptors in El País for news about “amnesty” over 25 years]”. *Infonomy*, 2(1), e24009.
*https://doi.org/10.3145/infonomy.24.009*



**Tomás Saorín-Pérez**
*https://orcid.org/0000-0001-9448-0866*
*https://www.directorioexit.info/ficha1039*
*Universidad de Murcia, Facultad de Comunicación y Documentación*
Campus de Espinardo, Edificio 3.
30071 Murcia, España
*tsp@um.es*

**Juan-Antonio Pastor-Sánchez**
*https://orcid.org/0000-0002-1677-1059*
*https://www.directorioexit.info/ficha1964*
*Universidad de Murcia, Facultad de Comunicación y Documentación*
Campus de Espinardo, Edificio 3.
30071 Murcia, España
*pastor@um.es*

**Resumen**

Este trabajo es una primera exploración sobre cómo ha evolucionado el uso de las palabras clave relacionadas con "amnistía" en el periódico *El País* a lo largo de 25 años. En lugar de usar el texto completo en lenguaje natural, toma como punto de partida la indización temática de las noticias ofrecida por el propio medio en forma de keywords de un vocabulario controlado. Una exploración convencional de los datos abre numerosas preguntas sobre el interés de los lenguajes controlados en la documentación periodística y cómo pueden servir como datos para investigaciones en humanidades digitales e information science, ofreciendo una nueva perspectiva sobre la construcción de significados y el análisis del cambio de uso y sentido de los términos.

**Palabras clave**

Palabras clave; Indización temática; Terminología socio-política; Evolución de significados; Vocabularios controlados; Temática de las noticias; Semántica.

**Abstract**

This work is a preliminary exploration into how the use of keywords related to "amnesty" in the newspaper *El País* has evolved over 25 years. Instead of utilizing the full text in natural language, it begins with the thematic indexing of news provided by the media itself in the form of keywords assigned from its controlled vocabulary. A conventional exploration of the data raises numerous questions about the interest of controlled languages in journalistic documentation and how they can serve as data for research in digital humanities and information science, offering a new perspective on the construction of meanings and the analysis of the change in use and sense of terms.

**Abstract**

Keywords; Subject indexing; Socio-political terminology; Evolution of meaning; Controlled vocabularies; News subjects; Semantics.

## 1. Introducción

Nos hemos acostumbrado a, como si tal cosa, procesar masivamente el lenguaje natural (PNL) de los documentos, y parece que cualquier otro enfoque es poca cosa. Planteamos aquí un caso de una investigación en curso sobre *keywords* usadas para indizar temáticamente noticias de actualidad. Las investigaciones en PLN trabajan con corpus textuales, para entrenar sus modelos con grandes datos de lenguaje real. ¿Qué pasaría si tratásemos al vocabulario controlado como un corpus textual?

Un vocabulario controlado es un lenguaje artificial construido para representar el contenido temático para su organización y acceso. Sin embargo, no deja de ser un lenguaje que presenta una pragmática de uso.

Esta pragmática es el acto de la indización. Y en consecuencia la asignación de palabras clave es un caso de uso tangible de un vocabulario que puede observarse desde distintos puntos de vista, como el cambio a lo largo del tiempo.

Aunque existen muchas fuentes para observar el uso de descriptores temáticos, se plantea un caso en el que tanto el vocabulario en sí mismo, como su aplicación en la indización, están sometidos a una relevante componente temporal, las noticias de prensa. La actualidad es un territorio que presenta unas cualidades singulares para observar la indización, puesto que está obligada a describir diariamente los acontecimientos cambiantes que construyen la actualidad diaria. Posee continuidad diaria, continuidad temática, continuidad de volumen de datos, y su observación a lo largo del tiempo plantea sugerentes preguntas sobre continuidad y cambio en el sentido, significado o uso de los términos controlados. Sería de esperar, además, que el cambio en las *keywords controladas* fuera parecido al cambio en el lenguaje de las noticias, en las formas de entender el mundo y en las categorías usadas para describir la realidad social.

Hay pocos datos explotables y públicos para estudios la evolución de la indización y el uso de descriptores en bibliotecas o bases de datos de todo tipo

Para desarrollar esta idea se necesita un buen conjunto de datos de noticias indexadas. Para ello se ha tomado como referencia al periódico *El País*, cuyas noticias se publican con una serie de keywords que permiten navegar por temas, personas, entidades o acontecimientos. Se trata de un conjunto de descriptores gestionado y coordinado por un departamento especializado en documentación, con suficiente rigor y continuidad para ofrecer palabras clave para noticias durante un periodo amplio. *El País* cuenta con un largo recorrido en un vocabulario colaborativo (**Rubio-Lacoba**, 2012), que está integrado por más de 130.000 términos de indización organizados en áreas, entre las que destacan: temas, personajes, organizaciones, lugares, y eventos (**García-Jiménez**; **Rodríguez-Mateos**; **Catalina-García**, 2019). Hay, por supuesto, otras muchas cabeceras nacionales o internacionales que tomar como fuente, pero, como veremos más adelante, no todas tienen las mismas políticas de indexación o de acceso a sus datos.

Por lo tanto, vamos a hacer una exploración preliminar de 25 años de noticias relacionadas con un tema explosivo que llena el espacio de debate público actual, la **amnistía**, y de camino, explicaremos el proceso de obtención de los datos.

## 2. ¿Qué nos puede decir el uso de un vocabulario controlado sobre los significados de amnistía?

En primer lugar, debemos detenernos, aunque sea muy brevemente, en el concepto de mismo de "keyword". Tomamos las siguientes definiciones de los glosarios del libro "Subject access to information" (**Golub**, 2014). Una acepción del término la entiende como

> "A word in natural language selected as representative of the content of an information resource, without any particular formal control".

Así pueden entenderse las keywords aportadas por los autores que se incorporan habitualmente en las publicaciones científicas. Un sentido parecido se puede encontrar, por ejemplo, en el campo del posicionamiento web, donde la investigación en

keywords (*keyword research*) sirve de soporte para conocer mejor los intereses de los usuarios en buscadores.

Sin embargo, en el campo *Library and Information Science*, este término "keyword" puede ser entendido como equivalente a "descriptor", es decir, término controlado que forma parte de un "Indexing language", es decir:

> "A specific type of controlled vocabulary representing formalized languages designed and used to describe the subject content of information resources for information-retrieval purposes"(**Golub**, 2014).

Estos vocabularios –tesauros, listas– están formados por términos precisos, compuestos por varias palabras conectadas que forman una unidad, o "término controlado", tanto en su forma preferida como en la no aceptada.

Por otra parte, muchos sistemas de información, y especialmente las plataformas sociales, usan "tags", que se entienden como

> "A keyword added freely by an author or user of a networked service to index a resource" (**Golub**, 2014).

Las keywords de *El País* las podemos entender como un vocabulario controlado (términos de indización), organizados en una amplia lista de términos aceptados (vocabulario controlado) que además requiere una alta dosis de renovación para reflejar los acontecimientos, por lo que funciona en una dinámica que demanda la creación de nuevos términos con mucha frecuencia, en donde colaboran redactores y documentalistas. Estas etiquetas (tags) además son navegables en la web e incorporadas como metadatos para los motores de búsqueda (*El País*, 2017).

Estudiar la secuencia de indización diaria de noticias sugiere formas de explicar la evolución del uso de ciertos términos clave de la actualidad

A pesar de estar inmersos "la primavera del lenguaje natural", el uso de lenguajes controlados para clasificar, indizar y anotar documentos sigue aportando una especificidad interesante a la que conviene prestar atención desde la óptica del análisis de datos. En humanidades digitales está cobrando fuerza la corriente "Collections as data", y queremos indagar en este mismo sentido en algo así como "Keywords as data".

### 3. Cambio (y resistencia) en los vocabularios controlados

El lenguaje, en su constante evolución, refleja y modela la realidad social y cultural de sus hablantes. No se trata de un fenómeno exclusivamente lingüístico, aunque se manifieste a través del lenguaje. Numerosas disciplinas estudian su evolución desde diferentes marcos teóricos y con diferentes propósitos, como la sociología lingüística o la historia de las ideas. Este trabajo se centra en los cada vez más numerosos campos que aplican "data-driven methods" para entender el cambio de las palabras, usos y sus significados. El estudio de la producción de significados presenta grandes complejidades, especialmente debido a que

> "some semantic changes occur in clusters, with a change in one word triggering a change in another" y, especialmente a que el cambio semántico "is much more closely connected with change in the external, non-linguistic world, especially with developments in the spheres of culture and technology" (**Durkin**, 2009, pp. 222-223).

El discurso de los medios de comunicación es un espacio especialmente favorable para este estudio, puesto que se combina en unidades fácilmente observables la realidad lingüística y la realidad social.

Las dinámicas de desajuste de significados que activan el cambio en las palabras o expresiones no solo se manifiestan en el lenguaje natural, sino también en los lenguajes artificiales diseñados para categorizar y organizar la información. Este trabajo aborda cómo el cambio semántico afecta sobre todo al terreno especializado de la indización de noticias periodísticas y el uso específico de la palabra clave "amnistía". La variación lingüística tiene muchas facetas. Denominaciones como "Evolving semantics" o "Semantic change" se encuentran en estudios de diversa metodología sobre la evolución de significados. También es un área en crecimiento aplicada a la investigación que observa y mide el fenómeno del cambio en el significado de los conceptos dentro de los modelos de representación del conocimiento: Clasificaciones, Tesuros, y Ontologías.

A partir de las metodologías de análisis de co-words se puede intentar explorar medir la continuidad o cambio en ciertos términos

## 4. Significados a la deriva

Puede entenderse la deriva semántica o "semantic drift" como la caracterización de desplazamientos de significado –en un sentido amplio– de términos en un determinado contexto de uso. Uno de los marcos metodológicos característicos de estos estudios es el "keyness analysis", que puede orientarse tanto a medir la diferencia como la similaridad (**Firoozeh** *et al.* 2019). En el campo del estudio de corpus lingüísticos se usa el término keyword para reflejar aquellos términos que caracterizan a un texto frente al otros. Este trabajo parte de un conjunto de datos en los que previamente se asigna una keyword en función de su utilidad para conectar y organizar noticias. El marco de estudio no es el de los estudios lingüísticos ni las tecnologías de procesamiento del lenguaje natural, sino el de los instrumentos para la organización del conocimiento en la tradición de la Library and Información Science, especialmente en la corriente denominada *Domain Analysis* (**Bawden**; **Robinson**, 2022; **Smiraglia**, 2016; **Gnoli**, 2020).

Además del "semantic drift", existen términos relacionados pero con significados ligeramente diferentes, para los cuales este estudio sirve de preámbulo o disparador de interés, según se observe principalmente el cambio, el desplazamiento, la pérdida de riqueza semántica y otros aspectos que admiten mayores precisiones formales (*Change*, *Shift*, *Decay*). El "seman-

El contenido de las noticias de actualidad pone a prueba la capacidad de los vocabularios controlados para descontrolarse

tic drift", por tanto, abarca la capacidad de los conceptos de ser reinterpretados por diferentes comunidades de usuarios o en diferentes contextos, introduciendo el riesgo de que pierdan su poder retórico, descriptivo y aplicativo. Este fenómeno puede clasificarse como colectivo no consistente, inconsistente o colectivo consistente, dependiendo de si el cambio es principalmente extrínseco o intrínseco y de cómo afecta las relaciones del concepto con otros.

Este análisis dual del lenguaje natural y artificial ofrece una comprensión más profunda de cómo la deriva semántica afecta la indización y representación de información en medios periodísticos, con un énfasis particular en el caso de estudio del diario *El País* y la evolución de la palabra "amnistía" presente en descriptores a lo largo de 25 años.

Los vocabularios –aka "lenguajes documentales"- evolucionan lentamente en sus términos y estructura. Son construcciones con tendencia a consolidar conocimiento, y no muy bien afinadas para situaciones de respuesta rápida. Sin embargo, un vocabulario de prensa posiblemente sea el caso de uso más flexible. Su uso diario en un contexto comunicativo real, realizado por los departamentos de documentación al indizar y organizar sus contenidos, requiere una adaptación continua de los temas de interés, modas y corrientes.

## 5. Amnistía, esa palabra

Podemos observar con precisión también el nacimiento, vida, muerte y resurección de las palabras clave

¿Qué sabemos de las palabras? Desde luego, muchas cosas, y el campo de la historia social de los conceptos es apasionante. Pero queremos partir aquí de los términos que se usan en instrumentos para organizar el conocimiento, puesto que las categorías y las etiquetas constituyen otro lenguaje, artificial y funcional, que delimitan y empujan el acceso a la información. A día de hoy, amnistía es una keyword relevante, pero ¿ha sido siempre así? Igual que confinamiento no significa lo mismo desde 2020, aunque Quevedo o Unamuno fueran confinados. Podría ser que un término que hoy aparece gritando y en negrita, hace unos años fuera uno más cualquiera entre tantos. Proyectos como el "Keywords Project" (*https://keywords.pitt.edu*) que registran y analizan

> "'keywords prominently used but also contested in social debate in English"

son de gran interés. Permiten un mayor ritmo que el alcanzado por monografías como "Keywords for today" (2018), inspiradas en el clásico de Raymond Williams "Keywords: A vocabulary of culture and society" (1976) o el "Diccionario político y social del siglo XX español" (**Aguilar-Fernández**, 2008).

La amnistía es un acto legislativo excepcional vinculado a episodios de cambio de ciclo en un conflicto sociopolítico. Las amnistías en España tienen bastante historia, casi siempre controvertida, y generan un grandísimo debate en la esfera pública mientras se fraguan, al materializarse y con posterioridad el valorarse sus efectos y defectos.

Estamos viviendo una controversia muy radicalizada ante la posible aprobación por el Parlamento español de una nueva ley de amnistía, en este caso dirigida a quienes par-

ticiparon en la declaración unilateral de independencia de Cataluña en 2017. Nuestra finalidad no es analizar las implicaciones jurídicas o sociales de la ley, sino tomarla como ejemplo de palabras y significados "agitados" y tratar de ofrecer ciertos datos para entender la actualidad desde la hemeroteca. Las palabras se las lleva el viento, y más las de la prensa. Así que, con Shakespeare de nuestro lado, "dejemos hablar al viento", y ya que leerse la hemeroteca del siglo lleva mucho esfuerzo, vamos a hacer unos snacks con los 25 años de #hashtags sobre amnistía(s).


Figura 1. Archivo de la transición. Carteles la legalización de los partidos y por la amnistía. Licencia CC BY
*https://archivodelatransicion.es/archivo-carteles/agrupados-por-motivos*

## 6. Dataset de keywords de *El País*: metodología de extracción

Este trabajo se basa en la obtención y explotación de datos de las noticias publicadas en la edición digital del diario *El País* durante el periodo 1999-2023. Cada noticia es indizada con un conjunto de palabras clave temáticas, que recogen tanto conceptos, como acontecimientos, personas o entidades. El análisis, por tanto, no se basa en el texto de las noticias, sino en el trabajo de indización temática de las mismas realizada por el departamento de documentación del periódico. Esta circunstancia implica trabajar sobre un vocabulario reducido y controlado, que crece conforme se demanda desde los propios acontecimientos noticiables. Este vocabulario es organizado y mantenido por un departamento especializado que trata de mantener la coherencia terminológica y su aplicación precisa en cada contenido publicado.
*https://www.um.es/metadatosvsdatos/la-amnistia-ya-no-es-lo-que-era-explorando-25-anos-de-keywords-en-las-noticias-de-el-pais*


Figura 2. Manifestación en Tarragona a favor de la Amnistía y el Estatuto de autonomía (1977). Foto: *Universitat de Barcelona*.
*https://www.counterfire.org/article/resisting-franco-the-assemblea-de-catalunya-50-years-on*


Figura 3. Pancarta por la libertad de los 'Jordis' en una manifestación, 19 nov. 2019. *Reusdigital.cat*.
*https://reusdigital.cat/noticies/catalunya/amnistia-veu-amb-preocupacio-les-condemnes-sedicio-i-reclama-la-llibertat*


Figura 4. Manifestantes piden amnistía para el músico catalán Pablo Hasél en Zaragoza. Fotografía de Christoph Pleininger, 2021. *Wikimedia Commons*. *https://commons.wikimedia.org/wiki/File:Protesters_calling_for_amnesty_for_the_Catalan_musician_Pablo_Has%C3%A9l_in_Zaragoza_-_1.jpg*


Figura 5. Manifestación contra la Amnistía, 10 de agosto de 2023, Barcelona, organizada por *Vox*. *https://commons.wikimedia.org/wiki/File:Manifestaci%C3%B3n_contra_la_Amnist%C3%ADa_08%C2%B710%C2%B72023_En_Barcelona_67.jpg*


Figura 6. Noticias online de *El País* mostrando las palabras clave aplicadas por su departamento de documentación.

## 7. Extracción y consolidación de datos

Otras cabeceras como *The New York Times* ofrece una portal de acceso a sus datos, que permite descargar mediante APIs sus artículos con gran nivel de detalle, incluyendo las palabras clave. En el caso de *El País*, las keywords se hacen claramente visibles para el lector en sus noticias digitales, pero hemos tenido que forzar su consulta con herramientas de rastreo web (*crawling*). Para la extracción de los datos de la investigación se ha usado el programa *Screaming Frog SEO Crawler*. Se realizó una análisis exploratorio de la estructura de páginas del diario *El País*, y se identificaron las páginas de archivo diario de noticias. Estas páginas usaban el patrón *https://elpais.com/hemeroteca/2016-01-30* y estaban organizadas en secuencias paginadas, por ejemplo *https://elpais.com/hemeroteca/2016-05-25/5*

El índice de noticias de un día cualquiera suele requerir entre 5/6 páginas de hemeroteca. Siguiendo este patrón se elaboraron listados con las url con todas las fechas de un semestre o año, que se usaron como input para el proceso de análisis y extracción de datos del sitio web. Se configuraron las opciones del módulo spider para recorrer solo *internal links*, *canonicals* y *pagination*. Además se activaron las opciones de "Check links outside start folder", "Crawl outside start folder" y "Crawl all subdomains", estableciéndose un límite de 8 niveles de profundidad y un máximo de 5 redirecciones. Los detalles de parametrización de la aplicación de rastreo los detallamos en nuestro blog de divulgación "Metadatos *vs.* Datos" como si fueran el clásico "anexo I". *https://www.um.es/metadatosvsdatos*

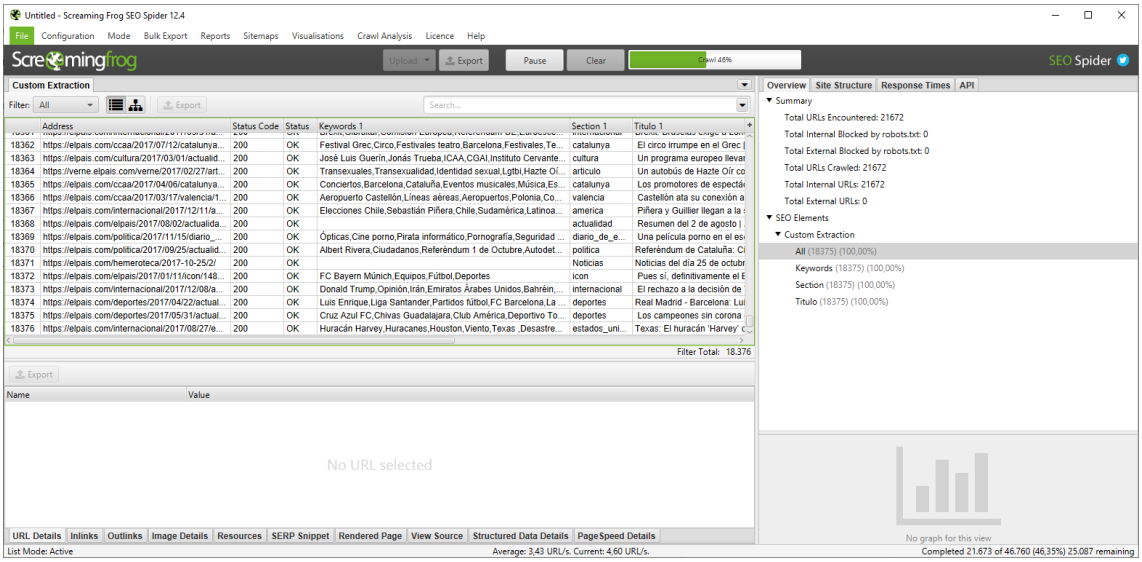


Figura 7. Ejemplo de la aplicación durante el proceso de extracción.

La principal dificultad fue delimitar con claridad el recorrido a partir de los enlaces facilitados, por lo que se configuró la herramienta para solo seguir enlaces dentro del dominio. Tras una exploración preliminar se observaron enlaces a contenidos complementarios y de servicio, así como enlaces relacionados con analítica y publicidad. Tras identificar sus patrones, se activaron criterios de inclusión y exclusión. También se controló que no hubiera duplicados usando las urls canónicas.

*Screaming Frog* es un programa de escritorio muy potente que también permite personalizar los datos que desean extraer (**custom_extraction**). En este trabajo apenas se usaron sus funciones orientadas a SEO, sino que se extrajeron elementos personalizados. Un estudio de las páginas en las que publicaban las noticias individuales, permitió identificar el código HTML apropiado. La url y el título no presentaban dificultades, pero para la obtención de las **keywords**, elemento esencial para esta investigación, no se pudo usar la etiqueta meta nativa de HTML, porque allí se presentaban en algo parecido a una secuencia de palabras en "lenguaje natural", quizá un vestigio de un sistema anterior de etiquetado automático de contenidos para SEO. Las palabras clave se localizaron declaradas individualmente mediante la *property meta* "article:tag". También se identificó una etiqueta personalizada **news_keyword**, presente en todas las páginas de noticias. En estos dos últimos casos, las keywords sí se correspondían con las visibles para el lector desde el navegador al final de los artículos, con función de enlaces implícitos de navegación por temáticas específicas relacionadas. Se consideró óptimo para la captura usar la serialización separada por comas.

La sección y la fecha no presentaban dificultades, puesto que existían etiquetas meta específicas para ello. Se parametrizó la extracción mediante Expresiones Regulares (Regex) o funciones XPath, para obtener los datos de las páginas de cada noticia.

Los datos obtenidos han requerido aplicar diversos procesos de consolidación y limpieza. Dado que el proceso depende de decisiones de publicación web de *El País*, se tuvieron que resolver algunas incidencias con la aparición, para algunos periodos de urls

extrañas o duplicación masiva de elementos, o construcción de urls que duplican contenido. El crawleo se ha realizado año a año, quedando pendiente la posterior unificación de todos los registros en un único dataset depurado de duplicados y de incidencias menores.

TODOS EN LA CALLE

La actriz de 'La casa de papel' Itziar Ituño se une a la manifestación por los presos de ETA en Bilbao

• La intérprete nacida en Basauri formó parte de la cabeza del evento, convocado por Sare

• Primeras palabras de Itziar Ituño tras la polémica por la manifestación por los presos de ETA: "No lo olvidaré nunca"

Itziar Ituño durante la manifestación (EFE)

LA VANGUARDIA
BARCELONA
15/01/2024 17:20 | Actualizado a 24/01/2024 10:20

Etiquetas • Instagram • Terrorismo • ETA • Manifestación
• Athletic Club de Bilbao • Audiencia Nacional
• Joseba Azkarraga • Policía Municipal

Figura 8. Este ejemplo de *La Vanguardia* sirve para ilustrar lo que sucede en los otros medios: Un porcentaje pequeñísimo de noticias tienen asignado "topics", y la gran mayoría tienen asignadas unas keywords automáticas que son apenas un recorte inanimado de palabras del titular.

Durante la extracción el rastreo del sitio web recuperaba un número considerable de páginas de índice o navegación, que no correspondían con una noticia individual (contenido) y que, por lo tanto, no tenían keywords temáticas.

Para todo el periodo analizado, de las páginas rastreadas, en el 88% de las páginas se obtenían resultados de la *custom extraction* planteada. No obstante, se detectan años con un comportamiento muy fiable en la captura de datos al rastrear el sitio web de *El*

*País*, con datos superiores al 90%, y un periodo entre 2012 y 2017 con porcentajes más bajos, entre el 54-77%. De estas páginas validas, el 73% contenían keywords. A pesar de estas dificultades de rastreo, los datos obtenidos son amplios y homogéneos entre todos los años analizados, con un promedio anual que superior a las 60.000 noticias. En total, el conjunto de datos incluye más 1.500.000 noticias con sus correspondientes descriptores.

Al hablar de la extracción hemos de señalar que el mismo intento de rastreo masivo sobre otros periódicos como *El Mundo*, *ABC* o *La Vanguardia*, presentaba resultados lamentables, porque su política de publicación de keywords en las noticias digitales es de mucha peor calidad (figura 8).

Una noticia con topic presenta señales de calidad de indización, por ejemplo "Instagram, Terrorismo, ETA, Manifestación, Athletic Club de Bilbao, Audiencia Nacional, Cárcel, Comunicación, Joseba Azkarraga, Policía Municipal", pero de una exploración masiva resulta que la mayoría presentan este aspecto: *<meta name="Keywords" content="actriz, casa, papel, itziar, ituño, une, manifestación, presos, eta, bilbao, mmn" />* Como apenas hay datos, no se puede realizar un análisis de largo recorrido temporal ni con suficiente cantidad de noticias relevantes.

## 8. Selección de los datos y construcción del dataset de amnistía

Una vez obtenidos los datos año a año, para los objetivos de este estudio se ha optado por extraer un subconjunto de noticias que usen una misma palabra clave. Se seleccionó un concepto en el que pueda observarse una dinámica intensa de cambio en su uso, debido a circunstancias de la actualidad. Al seleccionar el descriptor “Amnistía” se consideró que se enriquecería añadiendo también keywords que compartían el término principal, dado que eran un subconjunto muy pequeño y que no dificultaba los cálculos, procedimientos ni, sobre todo, mantener el foco explicativo en un caso único. Por lo tanto, se incluyeron también “Ley de amnistía”, “Amnistía fiscal”, “Amnistía internacional”, “Gestoras pro Amnistía” y “Amnistía 1977”, que son keywords de diferente significado y naturaleza.

Se asignó un identificador único a cada noticia (News-id) y se creó una tabla con las indizaciones individuales de cada noticia (indización), asignando un código único a cada Keyword (Kw-id). En este mismo proceso se normalizó y unificó la escritura de las palabras clave, de forma que, al procesar la serialización de las keywords separadas por comas, se eliminaron tildes, transformándolas a minúsculas y eliminando espacios en blanco al inicio o al final. Este mismo proceso de unificación de la grafía de las palabras clave se realizó también sobre las secciones.
Dataset:
*https://www.um.es/metadatosvsdatos/amnistia-25anos-keywords-elpais*

Este subconjunto de palabras clave permite adoptar una aproximación contextual al estudio del significado, aunque ha de tenerse en cuenta que es muy difícil separar entre significados y los “hechos enciclopédicos” que afectan de forma inevitable a las palabras: hoy es imposible leer amnistía no leerse igual que en 1977. Esta consideración nos sirve para resaltar que los datos seleccionados para nuestro análisis están

intensamente conectados al contexto, porque se produce desde la actualidad y con el fin de servir a presentarla y organizarla. Las keywords son un tipo de discurso funcional situado, caracterizable con tres elementos: Marco institucional (Medio periodístico); marco de referencia (actualidad nacional e internacional) y marco de producción (tratamiento documental).

El resultado final de la extracción nos aporta un conjunto de **2.284** páginas de noticias de *El País* de un periodo de 25 años, cada una de las cuales está indizada con varias keywords por el Servicio de Documentación del periódico, a partir de su propio vocabulario temático controlado. Esto supone un minúsculo porcentaje de todas las noticias de este período (0,141%). Las keywords tienen una distribución en subconjuntos relativamente separados, pues en pocos casos se indizan las noticias con varias de estas palabras clave al mismo tiempo.

Tabla 1. Las 6 keywords base del estudio

| **Keyword** | **Naturaleza** | **Nº de noticias** | **Fecha del primer uso** | **Identificador Kw-id** |
|---|---|---|---|---|
| Gestoras pro Amnistía | Organización | 171 | 4/01/2000 | 1982 |
| Amnistía internacional | Organización | 787 | 12/01/2000 | 356 |
| Amnistía | Concepto jurídico | 1046 | 21/10/2002 | 1832 |
| Amnistía 1977 | Acontecimiento político | 89 | 08/09/2008 | 2950 |
| Amnistía fiscal | Concepto específico | 238 | 30/03/2012 | 2866 |
| Ley de Amnistía | Acontecimiento político | 180 | 13/11/2023 | 2577 |

Las 6 keywords base del estudio (tabla 1) aparecen pocas veces usadas al mismo tiempo al indizar una noticia (aproximadamente en el 10% de los casos). Las noticias que finalmente estudiaremos tienen asignadas una media que ronda las 10 palabras clave.

## 9. Palabras (clave) en el tiempo

¿Qué nos dice una mirada temporal al uso de palabras clave? ¿Cuál era el "vocabulario de la amnistía" hace veinte o diez años? Veamos qué nos sugiere una de las formas más triviales de representar la convivencia de términos, una nube de palabras a partir de la frecuencia de uso de palabras clave en las noticias sobre amnistía en *El País* en cuatro momentos diferentes: 2001, 2012, 2019 y 2023 (figura 9): rápidamente identificamos episodios político-sociales que preocuparon a la sociedad española. Partiendo del terrorismo de *ETA* se ha ido pasando por los delitos de evasión de impuestos, al régimen penitenciario, los derechos humanos internacionales, hasta llegar a la explosión informativa de este año 2023 con los implicados en el *procés* independentista de Cataluña.

El volumen es otro indicador importante. El 38% de las noticias etiquetadas con alguno de estos descriptores de "amnistía" son de este último año, y más concretamente, casi del último semestre. Casi cualquier cambio de magnitud importante suele tener otras derivadas, y es otro aspecto a incluir en el análisis. Observamos además un etiquetado

donde cada vez cobran más importancia las personas frente a los conceptos, las instituciones o los acontecimientos.


Figura 9. Nubes de palabras clave más usadas en las noticias sobre amnistía en *El País* en cuatro momentos: 2001, 2012, 2019 y 2023.

Como curiosidad hemos realizado una *stop-motion* que en menos de 4 minutos pone a prueba nuestra agudeza visual y memoria, para identificar los conceptos que se barajan en el juego de la amnistía (figura 10).


**Nubes de Keywords en noticias sobre Amnistía(s) en 24 años de EL PAÍS**

Figura 10. *Stop-motion* con conceptos relacionados con amnistía. Versión de 24 años 2000-2023. *https://www.youtube.com/watch?v=if0fs-iTLgA*

El campo de bibliometría nos tiene acostumbrados a estudiar las redes temáticas, a través de cocitación, permitiendo entender algo mejor las comunidades de autores y las comunidades temáticas. Si aplicamos una clusterización convencional de co-words –en este caso co-keywords– usando *VOSviewer* a las noticias analizadas, obtenemos representaciones visuales más interesantes que con las nubes de palabras. Algunas de estas redes, como la de 2019, nos hablan de un sentido de las noticias sobre amnistía muy centrado en unos pocos términos: derechos humanos, amnistía internacional, solidaridad, ONGs. Estas keywords, si las entendemos como *named entities*, son a veces del tipo organización, a veces nombres de persona, y otras veces, lugares. También abundas los conceptos jurídicos y socio-políticos.

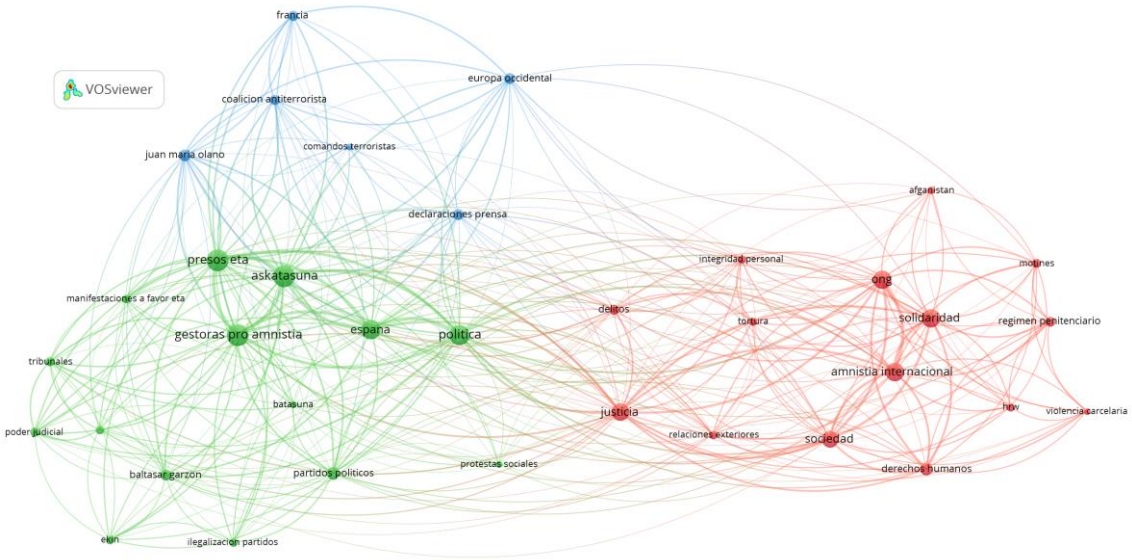



Gráfico 1. Keywords de amnistía, *El País*, año 2001.

Esta exploración preliminar nos hace pensar en que no sólo ha cambiado la realidad –han pasado cosas que remueven el universo temático de la amnistía– sino que también podría estar sucediendo cierto cambio de sentido al usarse. Esta evolución del significado, captada a través de términos de vocabularios controlados, es denominada a veces "deriva semántica" (*semantic drift*) porque los términos temáticos, al observarse en contexto en relación a otros, cambian de posición relativa, de relaciones, y podría servir para estudiar de otra manera la indización temática. La medición y representación de la deriva semántica de los conceptos no es una tarea sencilla, y puede ser afrontada con muchas toneladas de arsenal analítico. Sin embargo, nos ha parecido interesante compartir una visualización común de lectura fácil, las nubes de etiquetas, y una convencional en bibliometría, los clusters conectados. Partiendo aquí no del discurso primario, sino de una primera reducción de dimensionalidad a partir de la indización con palabras clave, podemos vislumbrar algunas ideas sugerentes. Todos podemos tener, de una u otra forma, una idea de cuáles han podido ser los diversos usos

del término, pero verlos representados de forma esquemática ayuda a releer nuestras ideas previas.

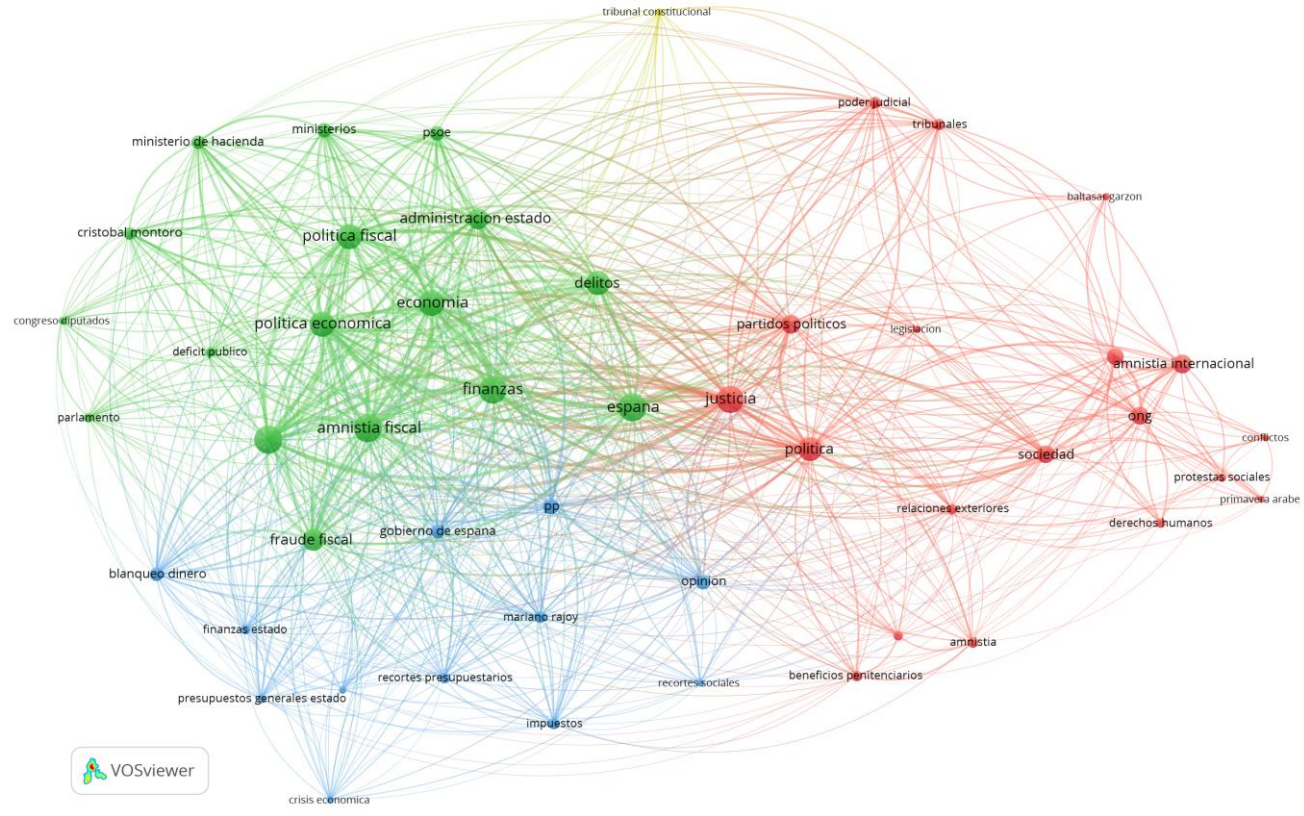



Gráfico 2. Keywords de amnistía, *El País*, año 2012.

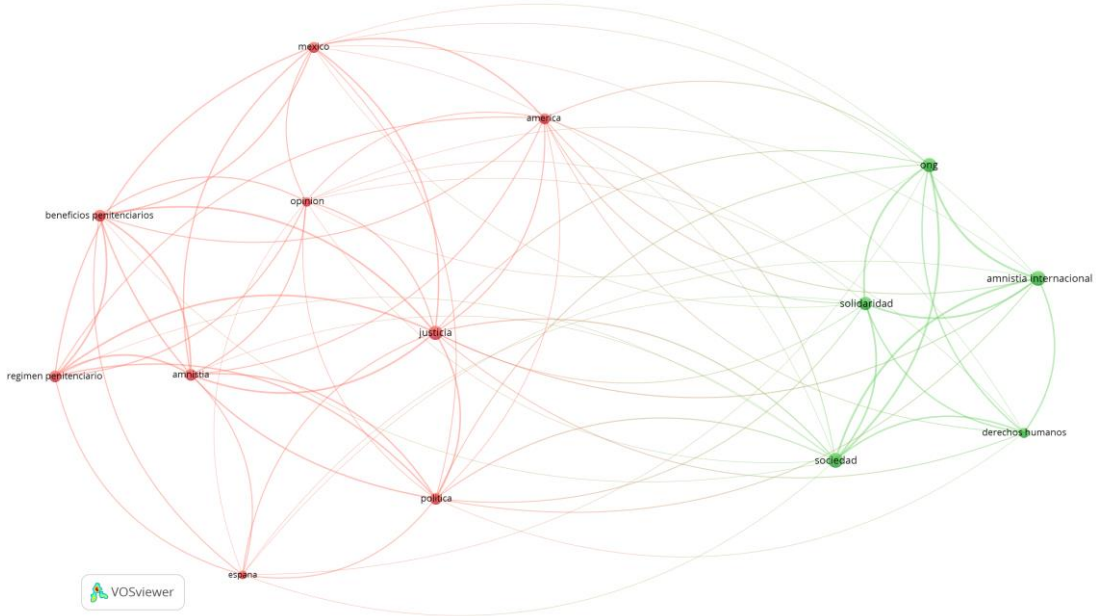



Gráfico 3. Keywords de amnistía, *El País*, año 2019.

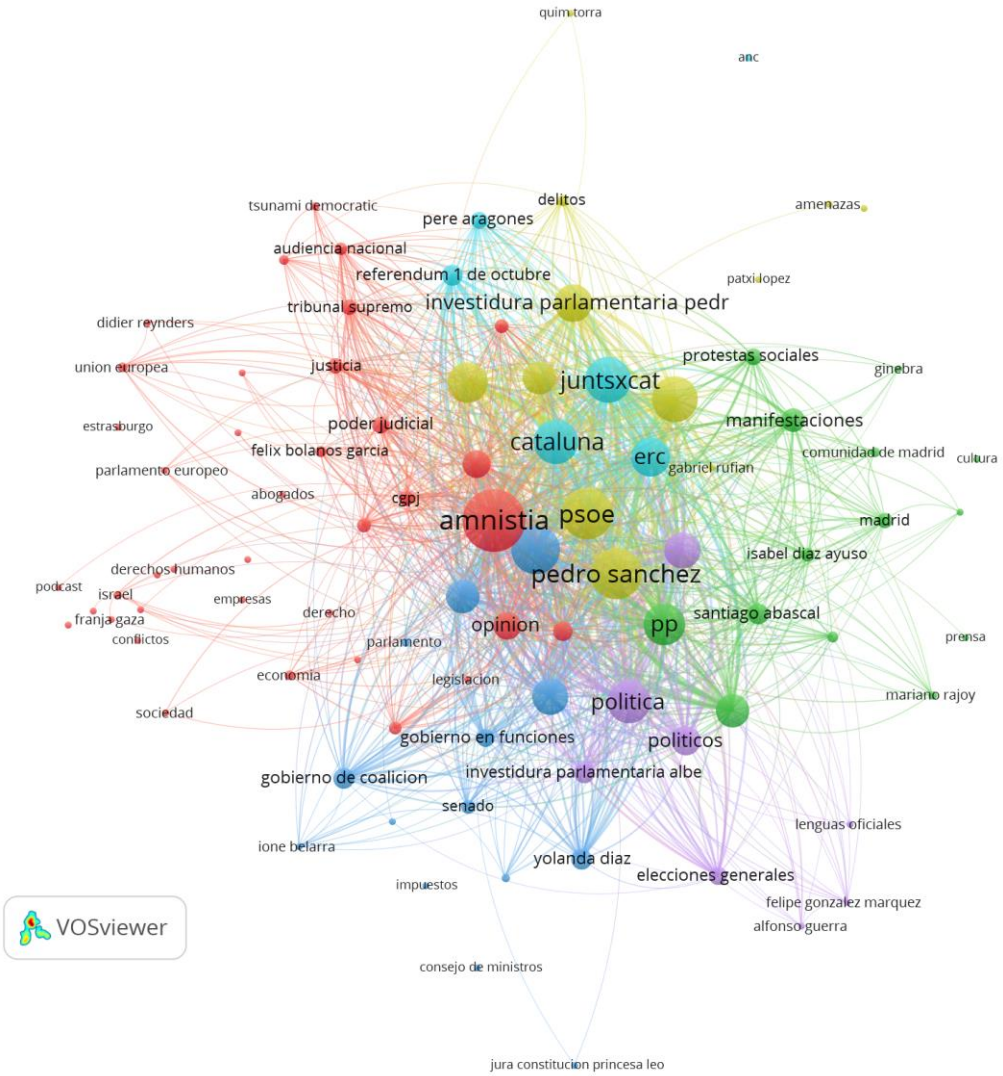



Gráfico 4. Keywords de amnistía, *El País*, año 2023.

Frente a los análisis de tendencias basados en *Google Trends*, que son de mucha utilidad, pero no abarcan con nitidez el contexto de uso, estamos explorando metodologías que permitan delimitar un universo completo en el que estudiar cuánto ocupa un tema en el registro diario de la actualidad, y sobre todo, con qué otros términos se enlaza y se presenta. Leer una hemeroteca así extractada, a base de palabras significativas seleccionadas para anotar las noticias, nos pone delante algunas ideas que van quedado relegadas por la agitación irascible de las noticias de política nacional: *delito fiscal*, *política penitenciaria*, *refugiados*, etc.

Existen muchas formas de aproximarnos al significado de un término a través de las palabras que le acompañan cuando se usa. Una de las formas más sencillas es un análi-

sis de **co-ocurrencias** de los términos de indización más frecuentes usados en las noticas seleccionadas (tabla 2).

Tabla 2. Diez keywords con mayor frecuencia de co-ocurrencia con los 6 descriptores controlados base de este estudio

| **Keyword** | **Top-10 co-ocurrencias Keywords** | **Keyword** | **Top-10 co-ocurrencias Keywords** |
|---|---|---|---|
| GESTORAS PRO AMNISTÍA | presos eta<br>askatasuna<br>espana<br>politica<br>justicia<br>tribunales<br>audiencia nacional<br>partidos politicos<br>juan maria olano<br>declaraciones prensa | AMNISTÍA INTERNACIONAL | ong<br>solidaridad<br>sociedad<br>justicia<br>derechos humanos<br>politica<br>delitos<br>espana<br>america<br>conflictos |
| AMNISTÍA | psoe<br>pedro sanchez<br>espana<br>politica<br>juntsxcat<br>investidura parlamentaria<br>cataluna<br>pp<br>carles puigdemont<br>erc | AMNISTÍA 1977 | amnistia<br>espana<br>politica<br>psoe<br>pedro sanchez<br>memoria historica<br>opinion<br>congreso diputados<br>justicia<br>cataluna |
| AMNISTÍA FISCAL | hacienda publica<br>politica fiscal<br>economia<br>politica economica<br>finanzas<br>justicia<br>espana<br>delitos<br>politica<br>administracion estado | LEY DE AMNISTÍA | Amnistía<br>España<br>PSOE<br>Pedro Sánchez<br>JuntsxCat<br>Congreso Diputados<br>PP<br>Política<br>Investidura parlamentaria Pedro Sánchez 2023<br>Investidura parlamentaria |

Esta tabla con las 10 keywords con mayor frecuencia de co-ocurrencia con los 6 descriptores controlados base de este estudio, permite ampliar los elementos con los que darle significado a cada palabra clave, aunque presenta varias limitaciones:

- No recoge la fuerza de la conexión de cada término con el término base. Es decir, no sabemos qué keyword tiene mayor frecuencia de aparición conjunta.
- No recoge la conexión interna entre estas palabras de extensión del contexto de uso. Es decir, no recoge el grado de co-ocurrencia entre varios términos, o los pares de elementos que co-ocurren con más frecuencia.
- Tampoco capta el sentido preciso con el que se ha usado al etiquetar noticias en cada momento. Es decir, no capta los términos que acompañaban y daban contexto al uso de una keyword en la fecha en la que sucedió la noticia.

- En línea con los principios clásicos de recuperación de información, se aprecia que ciertos términos muy frecuentes, como España, Política, Sociedad, no aportan valor para precisar el significado y podrían ser eliminados y entendidos de forma análoga a la sección (con la que, en muchas ocasiones, directamente se solapan).

Este último aspecto es especialmente relevante, puesto que los datos presentados anteriormente indican un desplazamiento/evolución de significados en el tiempo, según sean los acontecimientos sociopolíticos a nivel nacional e internacional.

Nos ha llamado también la atención que el lugar más estable para encontrar noticias sobre amnistía ha sido la sección de internacional, aunque en 2023 haya sido sepultada por el "tsunami amnistíaco".

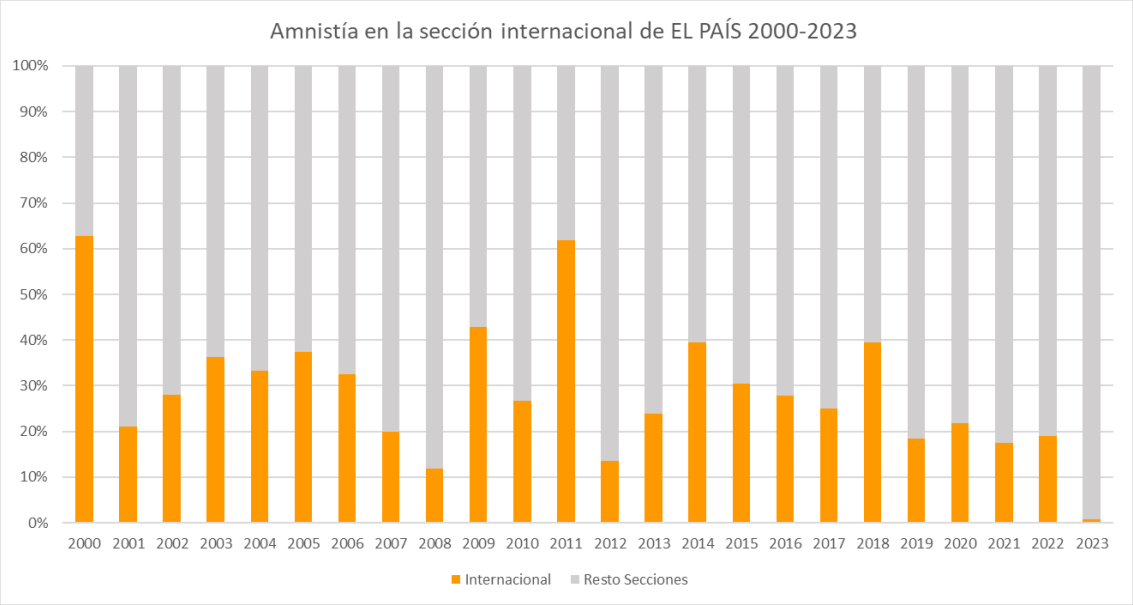



Gráfico 5. Porcentajes anuales de noticias sobre amnistía en la sección internacional de *El País*.

Igualmente, hemos observado que, igual que el 30 de marzo de 2012 los servicios de documentación de *El País* tuvieron que crear el descriptor "Amnistía fiscal" porque la actualidad había desbordado los términos que manejaban hasta ese momento, el 13 de noviembre de 2023 tuvieron que dar a luz el descriptor "Ley de amnistía". ¿Podríamos comparar la vida de ambos términos recién nacidos? Los descriptores parecen tener un ciclo de vida, como si fueran un virus oportunista que surge con fuerza, se expande, se diluye y, a veces, reaparece tímidamente con el paso de los años. Hemos jugado a transformar los datos para simular unos gráficos de calidad del sueño de estos descriptores cuando eran/son recién nacidos. En sus primeros pasos la *Ley de amnistía* apenas ha podido recuperar fuerzas con un buen sueño, mientras que la *Amnistía fiscal* pudo darse unas cuantas cabezadas.

Por otra parte, y ya para ir terminando, el gráfico 6 muestra el reparto de la tarta entre esos seis términos clave en competición.

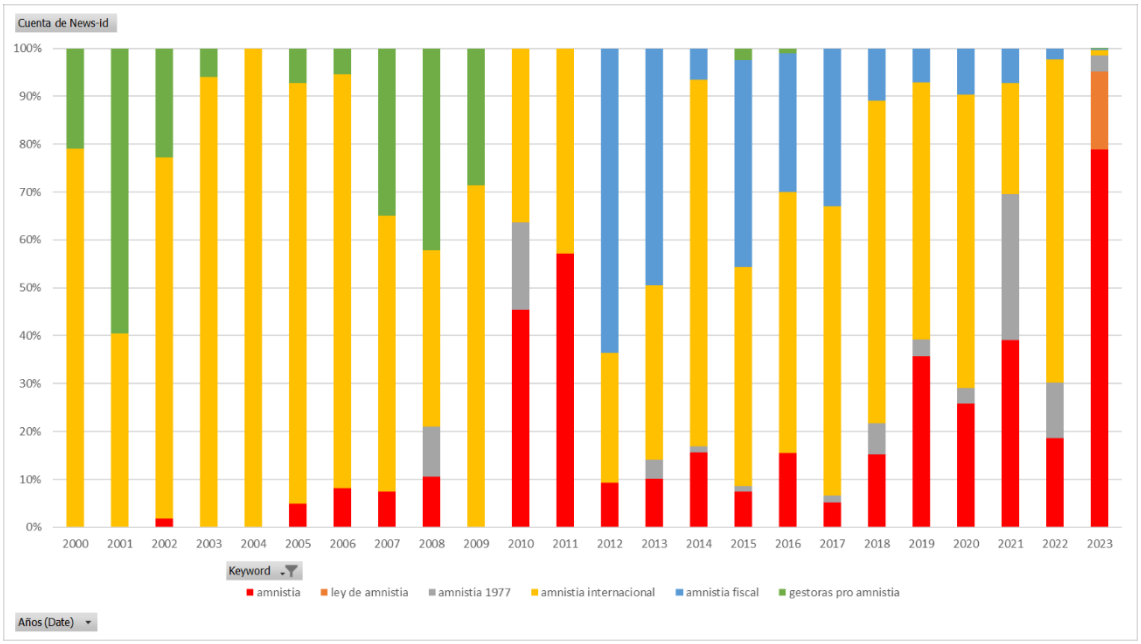


Gráfico 6. Porcentajes anuales de los 6 descriptores más usados que incluyen amnistía.

Estos seis descriptores son, además, un microcosmos interesante de observar: "Amnistía" necesita venir acompañada de otras palabras clave para que podamos entenderla. Igualmente, tomando distancia, la recién nacida "Ley de amnistía" quizá con el tiempo se llame "Amnistía 2024" igual que encontramos "Amnistía 1977" para referirse a una amnistía concreta. Además, encontramos dos autoridades cuasi-clásicas, la organización "Amnistía internacional" y "Gestoras pro amnistía", para agrupar a un conjunto amplio de colectivos que actúan en el mismo ámbito de interés. Dado que son términos que necesitan compañía para producir significado, encontrar mecanismos para representar estas conexiones, nos puede permitir comprender mejor la construcción de significados.

Hagamos notar, por último, que además de enfoques clásicos como los de Reinhart Koselleck en su "Historias de conceptos", o en la semántica histórica, el estudio de los instrumentos construidos para organizar el conocimiento –taxonomías, vocabularios, clasificaciones– también puede realizarse desde un punto de vista evolutivo, y sirve para construir sentido. ¿Cómo cambia nuestra forma de catalogar los objetos? Al tratar de representar y ordenar las cosas, aunque sea con discursos tan leves como los de las palabras clave, afectamos a lo observado, y nos costará conocer al mismo tiempo la velocidad y la posición.

Desde luego que podemos interpretar la realidad y explicar en qué sentido se usaba amnistía entre los años 2012 y 2014 y por qué fue necesario crear un nuevo término para indizar las noticias. Incluso hemos hecho la prueba de usar *ChatGPT* para la hipótesis interpretativa de partida y nos da una versión bastante acertada de la historia reciente de las amnistías en España (ver enlace a material complementario y anexos en la Nota 1), pero, podemos hacer que los datos nos cuenten esta misma historia, podemos capturar el momento en que una palabra empieza a virar su dirección principal

y adquirir nuevos sentidos. ¿Podríamos predecir en qué momento será necesario crear un nuevo término?

El "domain analysis", con referentes contemporáneos dentro del campo de la *Knowledge Organization* como Hjørland, Gnoli o Smiraglia, aporta metodologías e instrumentos para mapear un dominio, para organizar su información y proporcionar metadatos para manejar la marea desbordante de contenidos, datos, relatos y otras formas de la información cuya dinámica tiende siempre hacia un estado semisalvaje.

¿Podemos obtener una métrica que nos diga si un descriptor ha cambiado o mantenido su significado? ¿Podemos averiguar hacia dónde se ha desplazado? ¿Podemos describir por medios automáticos una aproximación a su significado en un momento dado? ¿Podemos estudiar el nacimiento, vida y muerte de las palabras clave? Existen numerosas preguntas de investigación que conectan los ámbitos del *topic modelling*, la indización automatizada, la clasificación automática o la agrupación de documentos, no solo con las taxonomías y vocabularios controlados, sino también con la actividad humana de indizar cuando selecciona los conceptos clave para representar temáticamente un contenido. En próximos trabajos indagaremos con más detenimientos y generalización a partir del conjunto de datos presentados aquí.

## Nota

1. Material complementario y datos en nuestra web de divulgación: "Metadatos vs. Datos: Exploraciones sobre organización de información digital, datos y contenidos" en *https://www.um.es/metadatosvsdatos*

## 10. Referencias


**Aguilar-Fernández, Paloma**. "Reconciliación". En: Fernández-Sebastián, Javier (coord). *Diccionario político y social del siglo XX español*. Madrid: Alianza, 2008, pp. 1024-1031.

**Bawden, David**; **Robinson, Lyn** (2022). *Introduction to information science*. London: Facet Publishing. ISBN: 978 1 783304950

*El País* (2017). "Así es el árbol del conocimiento de *El País*". *El país que hacemos*, 24 enero.
*https://blogs.elpais.com/que-hacemos/2017/01/así-es-el-árbol-del-conocimiento-de-el-país-.html*

**Firoozeh, Nazanin**; **Nazarenko, Adeline**; **Alizon, Fabrice**; **Daille, Béatrice** (2019). *Keyword extraction: Issues and methods*. Cambridge University Press.
*https://alvinntnu.github.io/NTNU_ENC2036_LECTURES/keyword-analysis.html*

**García-Jiménez, Antonio**; **Rodríguez-Mateos, David**; **Catalina-García, Beatriz** (2019). "Estudio sobre la indización/etiquetado y los lenguajes documentales en cinco diarios españoles". *Scire*, v. 25, n. 1, pp. 55-64.
*https://www.ibersid.eu/ojs/index.php/scire/article/view/4579*

**Gnoli, Claudio** (2020). Introduction to knowledge organization. London: Facet Publishing. ISBN: 978 1 783304677

**Golub, Koraljka** (2014) *Subject access to Information. An Interdisciplinary approach.* Bloomsbury Publishing, Libraries Unlimited. ISBN: 978 1 610695770

**Rubio-Lacoba, María** (2012). "Nuevas destrezas documentales para periodistas: el vocabulario colaborativo del diario *El País*". *Trípodos*, n. 31, pp. 65-78. *http://www.tripodos.com/index.php/Facultat_Comunicacio_Blanquerna/article/view/38*

**Smiraglia, Richard P.** (2015). *Domain analysis for knowledge organization*. Springer Books. ISBN: 978 0 081001509