SCIRE

Representación y organización del conocimiento

SCIRE

Representación y organización del conocimiento

Vol. 30, n.º 1, enero-junio 2024

ISSN 1135-3716

ISSN (e) 2340-7042

Scire:
knowledge representation and organization
Vol. 30, n. 1, January-June 2024

Ibersid:
Red de Investigación
en Sistemas de Información
y Documentación

ISSN: 1135-3716 = Scire (Zaragoza)
ISSN (e): 2340-7042
Depósito legal: Z. 1.790 — 1995

Edita: Ibersid® con la colaboración de Prensas de la Universdad de Zaragoza

Imprime:
Servicio de Publicaciones. Universidad de Zaragoza.
Edificio de Ciencias Geológicas, C/ Pedro Cerbuna, 12.
50009 Zaragoza, España. Tel.: 976 761 330. Fax: 976 761 063.

Scire

representación y organización del conocimiento

Alcance y objetivos

Scire: representación y Organización del Conocimiento es una publicación semestral de carácter interdisciplinar sobre la representación, normalización, tratamiento, recuperación y comunicación de la información y el conocimiento.

Difusión

Scire tiene difusión internacional. Agradecemos la inclusión en los siguientes servicios de referencia: Scopus, ESCI, Information Science Abstracts, Information Services in Physics, Electronics and Computing, Library and Information Science Abstracts, Sociological Abstracts, ERIH Plus, Knowledge Organization Literature, Base de Datos ISOC y Catálogo Latindex.

Instrucciones para los autores y procedimiento de evaluación

La última versión de las instrucciones para presentación de trabajos y del procedimiento de evalaución editorial están disponibles en: https://www.ibersid.eu/ojs/index.php/scire/about/submissions

Agradecimientos

Agradecemos el apoyo del Departamento de Ciencia, Universidad y Sociedad del Conocimiento del Gobierno de Aragón con su subvención a grupos de investigación T59_23R, al Vicerrectorado de Investigación y a la Facultad de Filosofía y Letras de la Universidad de Zaragoza.

Redacción, distribución y canje

Revista Scire
Departamento de Ciencias de la Documentación e Historia de la Ciencia
Facultad de Filosofía y Letras
Universidad de Zaragoza
C/ Pedro Cerbuna 12,
E-50.009 Zaragoza (Spain)

Tfno: int + 34 976 762239. Fax: 34 976761506.
E-mail: mailto:scire@unizar.es

Suscripciones y números sueltos

Suscripción anual: 30 €. Número suelto: 20 €. (IVA inc.)

Scire

knowledge organization and representation

Scope and aims

Scire: Representación y Organización del Conocimiento is an interdisciplinary journal published twice a year that is devoted to the representation, standarization, treatment, retrieval and communication of information and knowledge.

Dissemination

Scire has international distribution. We acknowledge its inclusion in the following reference services: Scopus, ESCI, Information Science Abstracts, Information Services in Physics, Electronics and Computing, Library and Information Science Abstracts, Sociological Abstracts, ERIH Plus, Knowledge Organization Literature, Base de Datos ISOC and Catálogo Latindex.

Instructions for authors and evaluation process

The last version of the instructions for authors and assessment process is available at: https://www.ibersid.eu/ojs/index.php/scire/about/submissions

Acknowledgments

We acknowledge the help of the Department of Science, University and Knowledge Society of the Government of Aragon (grant T59_23R to research groups), and of the Research Vice Rectorate and the Faculty of Philosophy and Arts of the University of Zaragoza.

Contact address

Revista Scire
Departamento de Ciencias de la Documentación e Historia de la Ciencia
Facultad de Filosofía y Letras
Universidad de Zaragoza
C/ Pedro Cerbuna 12,
E-50.009 Zaragoza (Spain)

Tel.: int + 34 976 762239. Fax: 34 976761506.
E-mail: scire@unizar.es

Subscriptions

Annual subscription: 30 €. Issue: 20 €. (VAT included)

Tabla de contenidos en español

Table of contents in Spanish

Tabla de contenidos en inglés

Table of contents in English

Análisis de Redes Sociales aplicado a las publicaciones científicas: el caso de la revista *El Profesional de la Información*

Social Network Analysis applied to scientific publications: the case of the journal Information Professional

José Luis ALONSO BERROCAL (1), Carlos G. FIGUEROLA (2)

Universidad de Salamanca, Instituto de Estudios de la Ciencia y la Tecnología, España (1) berrocal@usal.es. (2) figue@usal.es

Resumen

Se analiza, mediante el empleo de técnicas de análisis de redes sociales, la revista el Profesional de la Información (EPI). Se han analizado los años desde 2006 hasta 2023, mediante una descarga de los datos accesibles en el web, empleando un crawler diseñado ad-hoc. Los diferentes análisis de los datos, obtenidos a partir de los metadatos existentes, han permitido caracterizar las conexiones existentes entre los autores de las publicaciones, analizar la coocurrencia de palabras en los títulos de los artículos y detectar las comunidades existentes, tanto de autores como de palabras clave.

Palabras clave: Análisis de redes sociales. Detección de comunidades. Detección de tópicos. Revistas científicas. Profesional de la Información, El (revista).

Abstract

The journal El Profesional de la Información (EPI) is analysed using social network analysis techniques. The years from 2006 to 2023 were analysed by downloading the data available on the web using an ad hoc crawler. The different analyses of the data, obtained from the existing metadata, made it possible to characterise the existing links between the authors of the publications, to analyse the co-occurrence of words in the titles of the articles and to detect the existing communities, both of authors and of keywords.

Keywords: Social network analysis. Detection of communities. Topic detection. Scientific journals. Information professional (journal).

1. Introducción

Profesional de la Información es una de las revistas más importantes en el campo de la Ciencia de la Documentación (Library & Informaction Science) en español. Procedente de otra revista más antigua llamada Information World en Español (IWE), comienza como tal su andadura en 1998. Ha tenido una dilatada trayectoria, en la que su indización en las bases de datos más significativas es un hito importante.

Web of Science la indiza desde 2006 en los campos temáticos de Communication y en Information Science & Library Science:

> https://jcr.clarivate.com/jcr-jp/journal-profile?journal=PROF%20INFORM&year=2022)

En el primero de ellos aparece en Q3 en 2019 y en Q1 en 2022, mientras que en el segundo estaba en Q3 en 2019 y en Q2 en 2022.

Scopus, por su parte, la indiza también en ambos campos desde 2006:

> https://www.scopus.com/sourceid/6200180164

y la sitúa en 2022 en Q1 tanto con Communication como en Library & Information Science.

Se trata, en consecuencia, de una revista importante dentro de lo que podemos considerar la investigación en Ciencia de la Documentación reciente (entendiendo por reciente las dos últimas décadas, aproximadamente). El hecho de que se trate de una revista radicada geográficamente en España y, como se verá más adelante, de que sus autores sean mayoritariamente españoles y pertenecientes a organizaciones españolas, sugieren una especial representatividad de la investigación española de calidad en el campo de las Ciencias de la Documentación.

Así pues, tiene sentido analizar la producción publicada en esta revista desde diversos puntos de vista. Entre ellos, desde las autorías y las redes de colaboración; y también desde la estructura temática y su evolución. En lo que sigue, este trabajo se estructura de la siguiente manera: primero, se ofrece un estado del arte en las técnicas relacionadas con las coautorías y el análisis y modelado de temas aplicando técnicas de redes. Después se explicará la metodología seguida para la recolección y preparación de datos. A continuación, se expondrán y discutirán los resultados obtenidos, para finalizar con unas conclusiones.

2. Estado de la cuestión

2.1. Análisis de coautoría y de coocurrencia de palabras

Los investigadores nos tenemos que enfrentar a un entorno que cambia de forma muy rápida, debido al desarrollo constante de técnicas, herramientas y métodos científicos. Los niveles de colaboración entre los investigadores han aumentado de forma significativa, con la idea principal de mejorar tanto los niveles de calidad, como la eficiencia y la visibilidad de la investigación científica (Beaver, 2001).

Este crecimiento de las colaboraciones se produce en diferentes especialidades (Cronin, 2004), así tenemos que no solamente crecen en ciencias físicas (Wuchty et al., 2007), si no en ciencias sociales (Ossenblok et al., 2014). Un mayor detalle se puede encontrar en Henriksen (2016), con un detallado estudio del aumento en diferentes disciplinas de las ciencias sociales.

Una buena forma de demostrar las redes sociales que se generan entre investigadores, en un determinado campo, es el análisis de la coautoría. Este es un campo de estudio muy popular en investigaciones bibliométricas, y que abarca estudios de cocitas entre los autores, la coautoría y también análisis de coocurrencia de palabras (Leydesdorff & Vaughan, 2006). En el análisis de la coautoría podemos observar la evolución existente en las redes sociales entre los investigadores. Un aspecto importante de esta línea de trabajo es identificar los actores principales y sus relaciones (Perianes-Rodríguez et al., 2010).

Las redes se componen de un conjunto de actores (los autores de los artículos) y las interacciones que se producen entre ellos. Las interacciones las forman los lazos que relacionan a los autores y que permiten construir la estructura social. Los componentes principales de la red social son el actor y su vínculo relacional. En Análisis de Redes Sociales los actores de la red social se corresponden con el término nodo (extraído de la teoría de grafos). Las conexiones relacionales son las relaciones entre los actores, que incluyen diferentes tipologías de las relaciones, de colaboración, de amistad, flujos personales de una organización, etc. (Molina, 2001; Wasserman & Faust, 2013). Los lazos de red son los patrones relacionales estables que representan la posición de un individuo dentro de una red (Stevenson & Greenberg, 2000).

Los trabajos publicados sobre el análisis de redes sociales de coautoría son muy variados y con numerosos puntos de vista analizados. (Newman, 2004) trabajó con redes de coautoría en Biología, Física y Matemáticas. Los resultados reflejaron que la mayoría de los investigadores tienen solo unos pocos coautores, mientras que algunos tienen muchos (Adamic, 2000). Leydesdorff & Wagner (2008) trabajaron en la red de colaboraciones que se producía en artículos de Science Citation Index y llegaron a la conclusión de que la red de ciencia, en su conjunto, está aumentando.

Otros autores han trabajado en el análisis de las posiciones de los autores y cómo afecta a las redes de colaboración (Kong et al., 2019), y muchos trabajos que incorporan el cálculo de medidas propias del análisis de redes sociales, como intermediación, cercanía, diámetro, etc. para analizar desde otras perspectivas estas redes (Cheng et al., 2019).

2.2. Detección de temas emergentes

El Procesamiento del Lenguaje Natural (NLP) es una disciplina que nace en 1960 ligada a la Inteligencia Artificial y la Lingüística. Su principal objetivo es estudiar los problemas derivados de la generación y comprensión automática del lenguaje natural.

Gracias a las aportaciones realizadas en los últimos años en los métodos de trabajo, se pueden procesar enormes cantidades de información textual. En las áreas de Recuperación de Información Textual las técnicas de NLP son especialmente útiles.

En los sistemas de recuperación de información se emplea a menudo el procesamiento estadístico del lenguaje. A partir de este modelo cada documento está descrito por un conjunto de palabras clave que lo representan. Este concepto se conoce como bolsa de palabras (bag of words). Consiste en que las palabras de un documento se asignan como palabras claves con un peso en función de su importancia dentro del documento, basándose generalmente en la frecuencia de aparición. De esta manera, todas las palabras que tiene el documento se consideran como términos índices para ese documento. Además, se asigna un peso a cada término en función de su importancia, determinada normalmente por su frecuencia de aparición en el documento.

Una de las aplicaciones del NLP son las técnicas automáticas de clasificación de textos, que permiten asignar automáticamente categorías a una colección de documentos existentes. Uno de los puntos fundamentales de la aplicación de estas técnicas es la detección de topics, en inglés Topic Detection. Estos métodos son un conjunto de

algoritmos cuyo objetivo es descubrir los principales temas dentro de una colección estructurada de documentos. Por tanto, definir una estructura temática en una colección de documentos textuales. En los últimos tiempos estas técnicas se han aplicado a la web y a las redes sociales gracias a las nuevas herramientas tecnológicas (Blei, 2012). Debemos tener en cuenta que el Topic Detection and Tracking (TDT) es un track que comenzó en las Text Retrieval Conferences (TREC) (Allan, 2002).

Son métodos estadísticos que analizan las palabras de los textos originales para describir los temas que están tras cada documento. Además, permite organizar los archivos o resumirlos de forma automática. Para explicar el funcionamiento del modelo partimos de una colección de documentos en los cuales queremos identificar los temas subyacentes para poder organizar la colección. Cada documento contiene una mezcla de diferentes temas. Un tema es un conjunto de palabras que poseen una distinta probabilidad de aparición en los documentos. Por tanto, se asignan palabras al azar a cada tema y se realizan distintas pruebas hasta generar los mejores resultados y más consistentes (Underwood, 2012).

Para que el modelo sea efectivo es necesario probar distintas variaciones en la elección de palabras vacías o sobre el número de temas. Por tanto, el modelo se puede adaptar a la perspectiva o necesidades del investigador.

Uno de los algoritmos que mejor resultado ofrece es el LDA ("Latent Dirichlett allocation") diseñado por (Blei, 2012). Es un modelo estadístico que trata de captar la intuición para definir los temas que subyacen en cada documento de la colección que se analiza. Es por tanto un conjunto de algoritmos capaces de detectar y extraer las relaciones semánticas latentes en cada documento.

Se trata de un modelo de aprendizaje supervisado para caracterizar el contenido de los mensajes (Benhardus & Kalita, 2013). En él, cada documento se representa mediante una bolsa de términos, que son la única variable observada, a través de la cual se intenta extraer los temas tendencia. Es decir, LDA tiene como entrada un conjunto de términos correspondientes a la representación de cada documento de la colección, dando como salida los temas latentes de cada conjunto de palabras, que es lo mismo que los temas latentes de cada documento. Formalmente, un documento se asocia con una distribución multinomial de temas que a su vez son distribuciones multinomiales de palabras.

Un topic o tema es un conjunto de palabras que tienden a coocurrir juntas en los mismos contextos. Para ello los programas analizan su frecuencia de aparición en cada documento. El número de temas se define en un primer momento y puede ser reconsiderado en función de los resultados. Todos los documentos de la colección comparten el mismo conjunto de temas, pero cada documento pertenece a un tema de forma mayoritaria. No siempre este proceso es fácil y algunos documentos pueden presentar similitudes hacia varios temas diferentes. Por ello cada documento muestra la proporción que posee para estar relacionado con ese topic. El propio sistema es el que clasifica cada documento dentro de cada topic, una tarea manual que para una gran colección de documentos sería impensable.

Por tanto, el modelo LDA representa cada documento como una mezcla de los temas en los que aparecen ciertas palabras con sus probabilidades. El modelo LDA funciona de la siguiente manera (Cheng et al., 2019). Partimos de la base de un conjunto de documentos en el que hemos elegido un número fijo de temas T para descubrir. Deseamos conocer la representación de los temas en cada documento y las palabras asociadas a cada tema. Básicamente, el análisis de temas con LDA presupone que, en una colección de documentos hay n temas, representados por un conjunto de términos; cada documento contiene una determinada proporción de cada tema. Esa proporción puede ser 0 para determinados temas en determinados documentos, aunque lo más habitual es que se trate de cantidades muy pequeñas.

El modelo LDA nos proporciona una herramienta poderosa para el descubrimiento y explotación de la estructura temática de los archivos. Sin embargo, la formulación del LDA como modelo probabilístico presenta más ventajas y posibilidades. Una ampliación a este modelo es la incorporación de metadatos en los documentos tales como autor, títulos, enlaces, etc. Por ejemplo, en muchas colecciones de documentos estos están unidos por citaciones o relaciones entre las páginas webs mediante hipervínculos.

En este punto el modelo relacional de temas supone que cada documento se modela como en el LDA y los vínculos entre los documentos dependen de la distancia hacia las proporciones de los temas de cada documento. Es una buena solución para la gestión, organización o descripción de una gran cantidad de archivos de texto. Podemos decir que este método de análisis de temas puede ser aplicado en diversos campos como: ciencia política, bibliometría, medicina, piscología, etc. Debemos considerar que este tema es un nuevo campo emergente en el aprendizaje de

máquina en el que aún queda mucho por investigar en el futuro.

2.3. Detección de comunidades

En el contexto de las redes, al hablar de comunidad nos referimos a un conjunto de nodos de la red que están más densamente conectados entre sí que con el resto de la red. Existen muchas técnicas para la detección de comunidades, como los algoritmos de agrupamiento jerárquico, métodos basados en cliques, agrupamiento por cortes, algoritmo Girvan-Newman, etc.

Un método ampliamente utilizado es el análisis de modularidad (el número de vínculos entre grupos es pequeño, dentro de grupos es alto), destacando el algoritmo Louvain.

Para finalizar con este apartado queremos mencionar otros trabajos que analizan publicaciones de revistas, en diferentes ámbitos, en su mayoría trabajando sobre los resultados de WoS: el trabajo de Guerrero-Castillo et al. (2023), donde se realiza un análisis de la misma revista, centrándose en el estudio de las palabras clave y generando las comunidades que se obtienen a partir de la coocurrencia de la mismas; el trabajo de Alonso Berrocal (2022) en el que se analizan la autoría y la detección de topics de los artículos publicados en la revista *Fonseca Journal of Communication*, en la década 2010-2020; el trabajo de Guallar et al. (2020) realiza un estudio bibliométrico de 1.226 publicaciones entre los años 2015-2019, pertenecientes a las revistas españolas indexadas en WoS en la categoría Information Science & Library Science; otro trabajo que analiza el Profesional de la Información, en el periodo 2006-2017 es el de López-Robles et al. (2019), en el que se trabaja sobre 1.308 artículos extraídos de WoS y aplicando el software SciMAt; el trabajo de Olmeda-Gómez et al. (2017) analiza los temas de 2.247 artículos de autores españoles indexados en revistas de Información y Documentación en WoS en el período 1985-2014; otros estudios que analizan 580 artículos publicados en las siete revistas españolas indexadas en Scopus o WoS en el período 2012-2014 son (Ferran-Ferrer et al., 2017; Guallar et al., 2017).

La aportación del trabajo, frente a los más recientes realizados sobre la misma publicación, ofrecen una mecanismo de recogida de datos no basado en la descarga de datos sobre WoS (se realiza mediante un crawler), se utiliza el cálculo de índices de red globales, que caracterizan la red completa (diámetros, densidad): índices de centralidad (intermediación, pagerank) que caracterizan las características de cada nodo individualmente y se aplica un algoritmo de detección de comunidades, que entendemos es más eficaz con el mecanismo de recolección de datos realizado.Ofrece una visión desde una nueva perspectiva, dado que los estudios anteriores sobre la publicación utilizan procedimientos y software muy similar.

3. Metodología

La revista se publica en abierto en Internet, utilizando la plataforma Open Journal Systems (OJS) (Alperin et al., 2019) y se ha recogido la información que comprende los años 2006-2023, que en el momento de la recogida de datos suponía el 100% de los publicado. Como es habitual con este sistema, desde una página principal se accede a cada número publicado y, desde aquí, a los contenidos de cada número. Cada artículo, del tipo que sea, tiene su propia página.

La página de cada artículo lleva metadatos siguiendo la lógica establecida por el propio sistema OJS: Dublin Core (Weibel & Koch, 2000), por un lado y, por otro, metadatos para la indización por Google Scholar (Noruzi, 2005). Las Figuras 1 y 2 muestran algunos de los metadatos nativos de un artículo.

```
<meta name="gs_meta_revision" content="1.1"/>
<meta name="citation_journal_title" content="Profesional de la información
<meta name="citation_journal_abbrev" content="EPI"/>
<meta name="citation_issn" content="1699-2407"/>
<meta name="citation_author" content="Mariza Almeida"/>
<meta name="citation_author_institution" content="Universidad Federal del E
<meta name="citation_author" content="Igone Porto-Gómez  "/>
<meta name="citation_author_institution" content="Universidad del País Vasc
<meta name="citation_author" content="Loet Leydesdorff"/>
<meta name="citation_author_institution" content="University of Amsterdam"/
<meta name="citation_title" content="Are Brazilian innovation systems innov
<meta name="citation_language" content="en"/>
<meta name="citation_date" content="2023/12/29"/>
<meta name="citation_volume" content="32"/>
<meta name="citation_issue" content="7"/>
<meta name="citation_doi" content="10.3145/epi.2023.dic.07"/>
```

Figura 1. Metadatos Google Scholar

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Creator.PersonalName" content="Mariza Almeida"/>
<meta name="DC.Creator.PersonalName" content="Igone Porto-Gómez  "/>
<meta name="DC.Creator.PersonalName" content="Loet Leydesdorff"/>
<meta name="DC.Date.created" scheme="ISO8601" content="2023-12-29"/>
<meta name="DC.Date.dateSubmitted" scheme="ISO8601" content="2023-12-12"/>
<meta name="DC.Date.issued" scheme="ISO8601" content="2023-12-29"/>
<meta name="DC.Date.modified" scheme="ISO8601" content="2023-12-29"/>
<meta name="DC.Description" xml:lang="en" content="A knowledge-based economy
<meta name="DC.Description" xml:lang="es" content="A knowledge-based economy
<meta name="DC.Format" scheme="IMT" content="application/pdf"/>
<meta name="DC.Format" scheme="IMT" content="application/pdf"/>
<meta name="DC.Identifier" content="87536"/>
<meta name="DC.Identifier.DOI" content="10.3145/epi.2023.dic.07"/>
<meta name="DC.Identifier.URI" content="https://revista.profesionaldelainforr
<meta name="DC.Language" scheme="ISO639-1" content="en"/>
<meta name="DC.Rights" content="Derechos de autor 2023 Profesional de la inf
<meta name="DC.Rights" content="https://creativecommons.org/licenses/by/4.0",
<meta name="DC.Source" content="Profesional de la información / Information
<meta name="DC.Source.ISSN" content="1699-2407"/>
<meta name="DC.Source.Issue" content="7"/>
<meta name="DC.Source.Volume" content="32"/>
```

Figura 2. Metadatos Dublin Core

Se descargaron las páginas de los artículos y se extrajeron los metadatos de cada uno. No todos los campos son relevantes para el presente trabajo; pero uno especialmente importante es el

DC.Type.articleType. En la (Tabla I) se muestran los diferentes tipos contemplados a lo largo del período analizado y la cantidad de artículos para cada uno de ellos. Obviamente, no todos los tipos (Secciones de la revista, en realidad) se han mantenido en el tiempo. Algunos son muy específicos de un determinado período: Artículos de investigación Covid-19, por ejemplo.

Tipo	*N.º*
Agenda	28
Análisis	469
Artificial Intelligence	16
Artículo especial	2
Artículos de investigación	1091
Artículos de investigación Covid-19	49
Artículos de revisión	30
Editorial	114
Entrevistas	15
Especial sobre la revista Educación y biblioteca	3
Evaluadores	2
Humor académico	1
Indicadores	41
Informe técnico	6
Letters	14
Nota técnica	2
Nota de investigación / Research note	4
Patrocinado / Sponsored	1
Reseñas	49
Software	8
Software documental	20

Tabla I. Diferentes tipos de campos contemplados

Igualmente, no todas las secciones son relevantes para el presente trabajo; Agenda, Humor Académico, son ejemplos claros. Se han considerado relevantes aquéllos que disponían de título específico, palabras clave, resumen y referencias.

Después de una revisión manual exhaustiva, los tipos o secciones que hemos considerado relevantes se muestran en la (Tabla II). Éste ha sido el material básico para nuestro estudio.

En general, los metadatos descargados adolecen de algunas inconsistencias en la codificación de los caracteres, aun cuando el propio código HTML en que van insertados esos metadatos especifique claramente la utilización de UTF-8. Se aplicaron utilidades estándar de conversión y, en los casos en que esas utilidades no fueron eficaces, se aplicaron scripts ad-hoc con expresiones regulares para la detección y posterior sustitución de los caracteres con problemas de codificación.

Tipo	*N.º*
Análisis	469
Artificial Intelligence	16
Artículo especial	2
Artículos de investigación	1091
Artículos de investigación Covid-19	49
Artículos de revisión	30
Editorial	114
Especial sobre la revista Educación y biblioteca	3
Indicadores	41
Nota de investigación / Research note	4
Software	8
Software documental	20
TOTAL	1847

Tabla II. Tipos o secciones considerados relevantes

4. Resultados

La revista publica artículos en español y en inglés, y una pequeña parte en portugués la Figura 3 muestra la distribución temporal de ambas lenguas y puede apreciarse que los artículos se publican mayoritariamente en español, siendo la presencia de artículos en inglés testimonial hasta 2012. A partir de ese año la presencia de artículos en inglés crece lenta pero ininterrumpidamente, hasta llegar, actualmente, a la tercera parte de los que se publican.

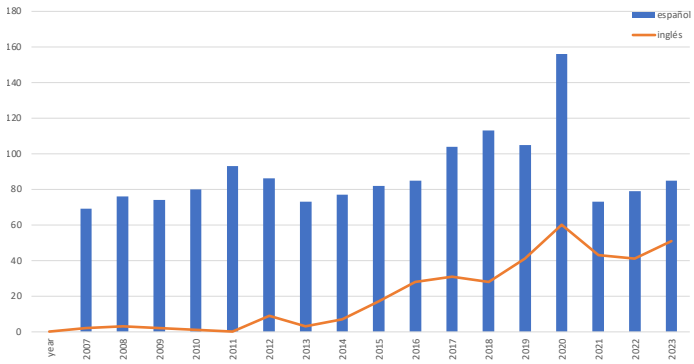

Figura 3. Distribución de idioma

Por lo que se refiere a los autores, es conocido el problema de la falta de homogeneidad formal en las autorías de la literatura académica (Medrano, 2020). A pesar de ello, y tras una revisión semimanual selectiva, se han podido identificar 2.385 autores únicos. La media de autores firmantes

por artículo es de 2,36 para todo el período; sin embargo, salvo momentos puntuales, el número de autores por artículo tiende a crecer, como puede verse la (Figura 4).

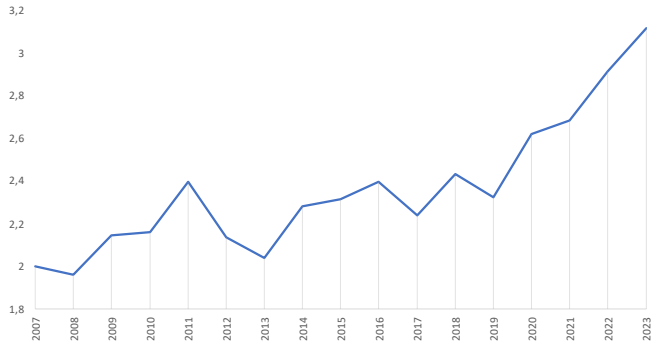

Figura 4. Promedio de autores por artículo

Autor	*Artículos*
Lluís Codina	44
Daniel Torres-Salinas	28
Javier Guallar	24
Félix De-Moya-Anegón	22
Emilio Delgado-López-Cózar	20
Blanca Rodríguez-Bravo	19
Ernest Abadal	18
Cristòfol Rovira	15
Mario Pérez-Montoro	15
Mike Thelwall	15
Xosé López-García	15
Andreu Casero-Ripollés	15
Carlos Arcila-Calderón	14
Antonia Ferrer-Sapena	13
Pere Masip	13
Enrique Orduña-Malea	13
Rafael Pedraza-Jiménez	13
Ramón Salaverría	13
Josep-Manuel Rodríguez-Gairín	13
Nicolás Robinson-García	13
Fernanda Peset	12
Loet Leydesdorff	12
Evaristo Jiménez-Contreras	11
Enrique Herrera-Viedma	11
Manuel Goyanes	11

Tabla III. Autores y frecuencia de publicación

De esos 2.385 autores, 1.600 firmaron un solo artículo en esta revista; 370 publicaron 2 artículos en todo el período estudiado; 25 autores han firmado más de 10 artículos. La Tabla III muestra los 25 autores que firman artículos en esta revista con mayor frecuencia. Estos autores, teniendo en cuenta el periodo de tiempo y los números publicados por la revista, forman lo que podríamos llamar el núcleo central de la revista, por lo que a autorías se refiere.

Como se ha comentado antes, pocos artículos se deben a un solo autor; la colaboración de varios conforma una red de autorías que permite percibir estructuras de colaboración, pero también temáticas, dado que los autores se especializan en campos temáticos muy específicos.

La red de coautoría permite ir más allá del simple número de artículos que firma cada autor. Así, podemos construir una red en la que los nodos son los autores, que conectan entre sí cuando comparten la autoría de un artículo.

La red resultante tiene 553 nodos o vértices (es decir, hay 553 autores que co-firman con al menos otro autor). Las métricas más relevantes de esa red se proporcionan en (Tabla IV).

Métrica	*Medida*
Grado medio	4,007
Densidad	0,007
Diámetro	24
Componentes conexos	36

Tabla IV. Medidas de red

Contrasta el nivel de grado con el diámetro de la red, considerablemente elevado, y más en una red de sólo 553 nodos. Esto sugiere una red estructurada en clusters muy autónomos, pero con sus nodos fuertemente interconectados entre sí, internamente a cada cluster; el número de componentes conexos apoya esta idea.

La propia silueta de esa red (Figura 5, en la página siguiente) muestra comunidades bien marcadas, con débiles conexiones entre unas y otras, al igual que pequeñas colaboraciones claramente periféricas.

La Tabla V muestra los 15 autores con mayor Page Rank y los 15 con mayor Intermediación (Freeman, 1977; Gleich, 2015); algunos de ellos son los mismos, aunque en diferente posición, salvo algunas excepciones. Una mejor posición en la tabla de Intermediación sugiere una mayor actividad multitemática, una mayor capacidad de colaborar con otros campos temáticos.

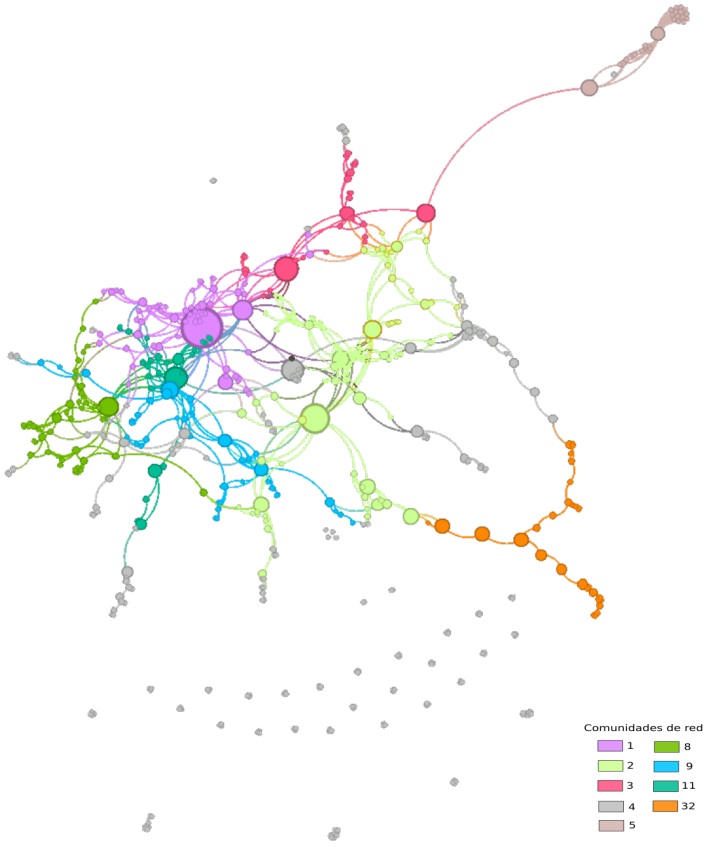

Figura 5. Visualización de comunidades de autores

Page Rank	Intermediación
Lluís Codina	Lluís Codina
Javier Guallar	Rafael Repiso
Daniel Torres-Salinas	Ernest Abadal
Xosé López-García	Ramón Salaverría
Félix De-Moya-Anegón	Félix De-Moya-Anegón
Ramón Salaverría	Javier Guallar
Antonio Castillo-Esparcia	Xosé López-García
Enrique Herrera-Viedma	Marta Somoza-Fernández
Javier Díaz-Noci	Enrique Orduña-Malea
Carlos Arcila-Calderón	Javier Díaz-Noci
Enrique Orduña-Malea	Daniel Torres-Salinas
Nicolás Robinson-García	Concepción Rodríguez-Parada
Ernest Abadal	Carmen Llorente-Barroso
Blanca Rodríguez-Bravo	Tamara Vázquez-Barrio
Pere Masip	Ruth Rodríguez-Martínez

Tabla V. Autores con mayor PageRank e Intermediación

En efecto, muchos autores se especializan en campos temáticos específicos, y esto tiene su reflejo en las comunidades de red. Si acudimos a la detección de comunidades de red (Plantié & Crampes, 2012) con los algoritmos habituales —Louvain o Leiden en nuestro caso (Traag et al., 2019)—, es fácil detectar comunidades de autores enfocados claramente en el mismo aspecto temático. La Tabla VI muestra las comunidades de autores más significativas.

N	*Autores*
1	Lluís Codina, Cristòfol Rovira, Guillermo López-García, Rafael Pedraza-Jiménez, Assumpció Huertas, Mario Pérez-Montoro, Santiago Tejedor-Calvo, Frederic Guerrero-Solé, Santiago Giraldo-Luque, Germán Llorca-Abad, Carles Pont-Sorribes, Ariadna Fernández-Planells, Isabel Villegas-Simón, Carlos Lopezosa, María Díez-Garrido
2	Rafael Repiso, Enrique Orduña-Malea, Daniel Torres-Salinas, Nicolás Robinson-García, Enrique Herrera-Viedma, Elea Giménez-Toledo, Adoración Merino-Arribas, Evaristo Jiménez-Contreras, Magdalena Trillo-Domínguez, Emilio Delgado-López-Cózar, María-Dolores Olvera-Lobo, Nuria Lloret-Romero, Isidro F. Aguillo, Luis Rodríguez-Yunta, Javier Salvador-Bruna
3	Ernest Abadal, Javier Guallar, Pere Masip, Núria Ferran-Ferrer, Julià Minguillón, Pere Franch, Manuel-Jesús Cobo-Martín, Juan-José Boté-Vericad, Remedios Melero, José-Ricardo López-Robles, Jaume Suau, Carlos Ruiz-Caballero, Sue Aran-Ramspott, Nadia-Karina Gamboa-Rosales, Lluís Anglada
4	Félix De-Moya-Anegón, Zaida Chinchilla-Rodríguez, Loet Leydesdorff, Lutz Bornmann, Benjamín Vargas-Quesada, Wenceslao Arroyo-Machado, Han-Woo Park, Rodrigo Sánchez-Jiménez, María-Victoria Nuño-Moral, Antonio Perianes-Rodríguez, Esteban Romero-Frías, Milan Martic, Veljko Jeremic, Vicente P. Guerrero-Bote, Carlos Olmeda-Gómez
5	Concepción Rodríguez-Parada, Blanca Rodríguez-Bravo, Ana-Reyes Pacios, Marina Vianello-Osti, Paz Fernández-y-Fernández-Cuesta, Andrés Fernández-Ramos, Marta De-la-Mano-González, David Nicholas, Eti Herman, Abdullah Abrizah, Chérifa Boukacem-Zeghmouri, Hamid R. Jamali, Jie Xu, Marzena Swigon, Carol Tenopir

N	Autores
6	Ruth Rodríguez-Martínez, Jesús Díaz-Campo, Salvador Gómez-García, María-Ángeles Chaparro-Domínguez, Amparo López-Meri, Xavier Ramon-Vegas, Salomé Berrocal-Gonzalo, Ana González-Neira, Andreu Casero-Ripollés, José-Luis Rojas-Torrijos, Marcel Mauri-Ríos, Teresa Piñeiro-Otero, Beatriz Feijoo-Fernández, Francisco Segado-Boj, Lluís Mas-Manchón
8	Xosé López-García, Carmen Costa-Sánchez, Francisco Campos-Freire, María-Isabel Míguez-González, José Rúas-Araújo, Valentín-Alejandro Martínez-Fernández, Miguel Túñez-López, Pablo Vázquez-Sande, María-José Establés, Olga Blasco-Blasco, Berta García-Orosa, María-José Ufarte-Ruiz, Vicente Coll-Serrano, Ana-María López-Cepeda, Marta Rodríguez-Castro
9	Javier Díaz-Noci, Carlos Arcila-Calderón, Ana Serrano-Tellería, Mar Iglesias-García, Elías Said-Hung, Ignacio Aguaded, Rosa Berganza, Lucía García-Carretero, Marta Martín-Llaguno, Luis-Mauricio Calvo-Rubio, Jordi Morales-i-Gras, Dolors Palau-Sampio, Félix Ortega-Mohedano, Sonia Parratt-Fernández, Ainara Larrondo-Ureta
10	Marta Somoza-Fernández, Alexandre López-Borrull, Josep-Manuel Rodríguez-Gairín, Roberto García, Tomás Saorín, Fernanda Peset, Mari-Carmen Marcos, Rafael Aleixandre-Benavent, Aurora González-Teruel, Toni Granollers-Saltiveri, Rosângela-Schwarz Rodrigues, Eva Ortoll, Candela Ollé-Castellà, Tomàs Baiget, Gregorio González-Alcaide
11	Ramón Salaverría, Josep-Lluís Micó-Sanz, Santiago Justel-Vázquez, José-Alberto García-Avilés, Pilar Sánchez-García, Montse Bonet, Bienvenido León, María-Carmen Erviti, Alicia De-Lara-González, Miguel Carvajal-Prieto, Gustavo Cardoso, Iván Lacasa-Mas, Toni Sellas-Güell, Alba Diez-Gracia, Félix Árias-Robles
12	Charo Sádaba-Chalezquer, Alfonso Vara-Miguel, Manuel Goyanes, Homero Gil de Zúñiga, Javier Serrano-Puche, Gloria Rosique-Cedillo, Alberto Ardèvol-Abreu, María-Pilar Martínez-Costa, Alejandro Barranquero-Carretero, Brigitte Huber, Florencia Claes, Alba Córdoba-Cabús, Luis Deltell, Carmen Rodríguez-Wangüemert, Eduardo-Francisco Rodríguez-Gómez
13	Antonio Castillo-Esparcia, Ana Castillo-Díaz, Iván Puentes-Rivera, Carmen Carretón-Ballester, Carmen Quiles-Soler, Miguel De-Aguilera-Moyano, Alba-María Martínez-Sala, Lucía Caro-Castaño, Ana-Belén Fernández-Souto, Ana Almansa-Martínez, Paula Pineda-Martínez, Concepción Campillo-Alhama, Inmaculada Berlanga-Fernández, Isabel Ruiz-Mora, Justo Villafañe-Gallego
14	Inmaculada J. Martínez-Martínez, Juan-Miguel Aguado-Terrón, Claudio Feijóo-González, José-Luis Gómez-Barroso, Sergio Ramos-Villaverde, Francisco-José Sarabia-Sánchez, Yannick Boeykens
15	José-Antonio Cordón-García, Raquel Gómez-Díaz, Julio Alonso-Arévalo, Helena Martín-Rodero, Javier Merchán-Sánchez-Jara, Almudena Mangas-Vega, Araceli García-Rodríguez
18	Beatriz Catalina-García, Esther Martínez-Pastor, Ricardo Vizcaíno-Laorga, Manuel Montes-Vozmediano, Joaquín López-del-Ramo, José-Antonio Merlo-Vega, Gema Alcolea-Díaz, Natalia Arroyo-Vázquez, José-Antonio Gómez-Hernández, Tony Hernández-Pérez, Pilar Beltrán-Orenes, David Rodríguez-Mateos, Miriam Rodríguez-Pallares, Aurora Cuevas-Cerveró, Manuel Fernández-Sande
19	María-Teresa Fernández-Bajón, José-Antonio Moreiro-González, Juan-Antonio Pastor-Sánchez, Diego Martín-Campo, Tránsito Ferreras-Fernández, Sonia Sánchez-Cuadrado, Jorge Morato-Lara, Carlos-Miguel Tejada-Artigas, María-Jesús Colmenero-Ruiz, Adilson-Luiz Pinto, Julián Urbano, Mónica Marrero, Audilio Gonzales-Aguilar, Miguel-Ángel Marzal-García-Quismondo, Valentín Moreno-Pelayo
24	Jordi Sánchez-Navarro, Leila Mohammadi, Pablo Lara-Navarra, Daniel Aranda, Silvia Martínez-Martínez, Ferran Lalueza, Elisenda Estanyol, Mireia Montaña-Blasco, Cristina Aced-Toledano, Judith Clares-Gavilán, Elena Neira, David Maniega-Legarda, Eva-Patricia Fernández-Manzano, Susana Miquel-Segarra
25	María-José Cantalapiedra-González, Íñigo Marauri-Castillo, Juan-José Gutiérrez-Cuesta, Leire Iturregui-Mardaras, María Ruiz-Aranguren
26	Natalia Papí-Gálvez, Alejandra Hernández-Ruiz, Juan-José Perona-Páez, María-Luz Barbeito-Veloso, Ana-María Enrique-Jiménez, Estrella Barrio-Fraile, Marta Perlado-Lamo-de-Espinosa, Sonia López-Berna
29	Antonio Hidalgo-Nuchera, Julián Chaparro-Peláez, Ángel Hernández-García, Santiago Iglesias-Pradas, Alberto Urueña-López, Antonio Fumero-Reverón
31	Emili Prado, Óscar Coromina, Celina Navarro, Matilde Delgado, Adrián Padilla, Belén Monclús, Núria García-Muñoz
32	Carmen Llorente-Barroso, Tamara Vázquez-Barrio, Rebeca Suárez-Álvarez, Mónica Viñarás-Abad, Francisco Cabezuelo-Lorenzo, María Sánchez-Valle, Javier Sierra-Sánchez, Francisco García-García, Vanessa Rodríguez-Breijo, Francisco-Javier Herrero-Gutiérrez, Daniel Barredo-Ibáñez, Núria Simelio-Solà, Pedro Molina-Rodríguez-Navas, Jorge Gallardo-Camacho, Luis Mañas-Viniegra
33	Sara Martínez-Cardama, Mercedes Caridad-Sebastián, Ana-María Morales-García, Ana R. Pacios, Fátima García-López
35	Concha Pérez-Curiel, Mar García-Gordillo, Ricardo Domínguez-García, Rubén Rivas-de-Roca
38	Jorge Clemente-Mediavilla, Nuria Villagra, Abel Monfort, Rebeca Antolín-Prieto

Tabla VI. Comunidades de autores ordenadas por intermediación

4.1. Instituciones de afiliación

De cara a establecer redes de colaboración, son importantes las instituciones a las que están afiliados los autores que colaboran en los diferentes artículos. Los metadatos de EPI reservan un campo para este dato; sin embargo, su utilización plantea algunos problemas.

El primero es que no todos los artículos ofrecen información al respecto; sólo 776 tienen información válida sobre esta cuestión. En el resto, en la mayor parte de los casos figura como única institución la expresión 'El Profesional de la Información' y, en otros, simplemente está vacío.

El segundo problema es el de la falta de uniformidad en la formulación de las Instituciones. Este problema es bien conocido (Kim & Diesner, 2015), y, de alguna manera, parecido al planteado por el de los nombres de los autores. Sin embargo, en este caso, carecemos todavía de mecanismos similares al ORCID (Haak et al., 2018).

Adicionalmente, tenemos la cuestión de los diferentes niveles de agregación: en unos casos, por ejemplo, figura como institución una universidad determinada, mientras en otros figura una facultad concreta de esa universidad, o incluso un departamento de una facultad de una universidad, y así sucesivamente.

Grado medio	*5,72*
Densidad	0,022
Diámetro	9
Componentes conexos	5

Tabla VII. Medidas de red para instituciones

Organizaciones con mayor grado
Universidad Complutense de Madrid
Universidad de Granada
Universitat Pompeu Fabra
Universidad de Málaga
Universidad Carlos III de Madrid
Universitat Autónoma de Barcelona
Universidad de Salamanca
Universitat Oberta de Catalunya
Universidade de Santiago de Compostela
Universitat de Barcelona
Universidad Rey Juan Carlos
Universidad de Navarra
Universidad de Valladolid
Universidad del País Vasco
Universidad Internacional de La Rioja

Tabla VIII. Instituciones ordenadas según su grado

Organizaciones con mayor Page Rank
Universidad Complutense de Madrid
Universidad de Granada
Universitat Pompeu Fabra
Universidad Carlos III de Madrid
Universitat de Barcelona
Universidad de Málaga
Universidad Rey Juan Carlos
Universidad de Salamanca
Universitat Autónoma de Barcelona
Universitat Oberta de Catalunya
Universidade de Santiago de Compostela
Universidad de Navarra
Universidad Internacional de La Rioja
Universidad del País Vasco
Universidad de Valladolid

Tabla IX. Instituciones ordenadas según su PageRank

Con todo, y tras una revisión manual de las instituciones más frecuentes, es posible obtener informaciones útiles. Así, de 548 instituciones diferentes, unificando manualmente los niveles de agregación, nos quedamos con 235 instituciones.

Construimos una red de coincidencias con las instituciones que colaboran (coocurren) en los mismos artículos, generando una pequeña red de 235 nodos y 596 enlaces. En la Tabla VII tenemos las medidas de red más relevantes y en las Tablas VIII y IX las instituciones ordenadas por valor de grado y de PageRank. En la Figura 6 podemos ver las redes de colaboración bien definidas.

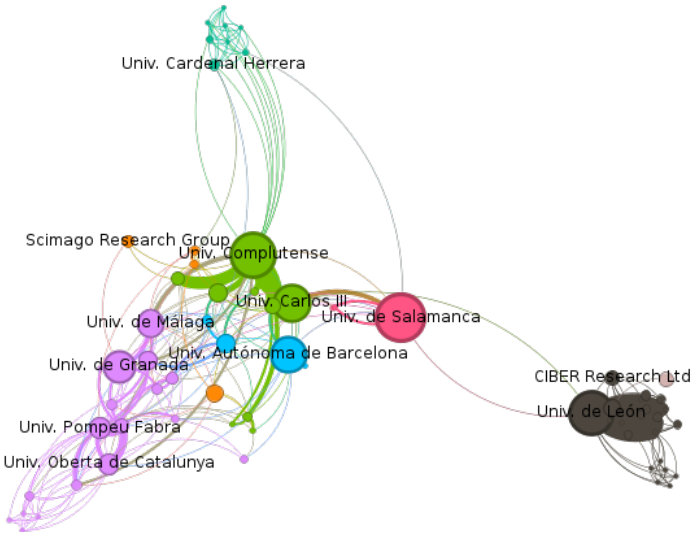

Figura 6. Red de colaboraciones entre instituciones

4.2. Palabras clave

Los metadatos de los artículos contienen palabras clave descriptivas de su contenido; se trata de palabras clave del autor o autores, por lo que no cabe esperar normalización ni uso consistente (Peset et al., 2020). En efecto, encontramos un total de 15.433 palabras clave (8,35 palabras clave por artículo, en promedio), y 6.019 palabras clave únicas.

Las palabras clave coocurren en determinados artículos, formando parejas coincidentes. De un total de 72.204 coincidencias, 62.457 suceden una sola vez. Así, se puede constatar una elevada dispersión en el uso de palabras clave, lo cual es consistente con lo observado en otros trabajos y lleva a plantear la utilidad de tales palabras clave.

Con palabras clave y sus coincidencias o coocurrencia en el mismo artículo, es posible también construir una red (no dirigida) en los que cada palabra clave es un nodo y la coocurrencia de dos de ellas constituye un arco o enlace entre dichos

nodos. La red resultante es marcadamente dispersa; pero con una simple poda de los enlaces que representan una única coocurrencia y la consiguiente de los nodos que resultan con grado igual a 0, resulta una red de sólo 1.334 nodos o palabras clave. Este enfoque ha sido aplicado en otros trabajos, como el de Lozano et al. (2019).

Grado medio	7.9
Grado medio con pesos	22.32
Diámetro	9
Componentes conexos	23

Tabla X. Medidas de red para palabras clave

espana, internet, comunicacion, periodismo, medios de comunicacion, periodismo digital, publicidad, medios digitales, prensa digital, diarios]
covid-19, social media, social networks, coronavirus, journalism, pandemias, disinformation, communication, pandemics, fake news
redes sociales, twitter, medios sociales, comunicacion politica, television, web 20, facebook, instagram, participacion, desinformacion
revistas cientificas, acceso abierto, scopus, bibliometria, produccion cientifica, web of science, bases de datos, repositorios, universidades, impacto
bibliotecas universitarias, bibliotecas, digitalizacion, bibliotecas publicas, indicadores, perfiles profesionales, documentacion audiovisual, evolucion, bibliotecarios, usuarios
big data, transparencia, sitios web, datos abiertos, marketing, usabilidad, webs, datos masivos, administracion publica, datos
radio, audio communication, podcasting, podcasts, audio, digital audio, europa, business models, podcast, platforms
spain, scholarly communication, open access, china, transparency, interviews, poland, france, research, early career researchers
relaciones publicas, comunicacion corporativa, comunicacion organizacional, gestion del conocimiento, empresas, profesionales, reputacion, comunicacion interna, responsabilidad social corporativa, organizaciones
web semantica, ontologias, wikipedia, sistemas de informacion, gestion documental, tesauros, clasificaciones, skos, vocabularios controlados, datos enlazados

Tabla XI. Palabras clave con mayor PageRank por comunidades

Algunos de los datos más relevantes de esa red aparecen en la Tabla X. El peso de los arcos se ha considerado como la simple frecuencia con que cada par de palabras clave coocurre. La formación de comunidades de palabras clave en esta red sugiere bloques de especialización temática, aunque el uso de palabras clave claramente transversales, por un lado; y otras de contenido muy amplio, por otro, desdibujan un tanto esas comunidades. La mezcla de palabras en inglés y español introduce también ruido.

Así, la Tabla XI muestra las palabras clave con mayor Page Rank de las comunidades más importantes obtenidas mediante el algoritmo de Leiden. En la Figura 7 (en la página siguiente) podemos observar la red formada por las palabras clave.

En cualquier caso, técnicas más sofisticadas de análisis de temas o topic modeling pueden ser aplicadas, dado que los propios metadatos incluyen los resúmenes de los artículos. El análisis automático de temas basado en los textos de los resúmenes ha sido aplicado en diversos trabajos de tipo bibliométrico, como en Figuerola et al. (2017) o en García-Marco et al. (2020) y se basa en la aplicación de técnicas de LDA (Latent Dirichlet Allocation).

Existen programas que realizan los cálculos e identifican los temas en los documentos de forma relativamente sencilla (McCallum, 2002).

En este trabajo se manejó un número inicial de 20 topics que, tras analizar los resultados, se refundieron o agruparon en 9. Como la técnica nos proporciona el porcentaje o proporción de cada tema en cada resumen de artículo y disponemos de la fecha de publicación de cada uno de esos artículos, es posible establecer la importancia de esos temas y su evolución temporal.

Los grandes temas presentes en la revista EPI son (Tabla XII):

Educación, formación especializada, universidad
Periodismo y medios de comunicación
Bibliotecas
Producción científica, bibliometría
Perspectiva de género
El libro y la edición, incluyendo libro electrónico
Redes sociales
Información y salud, incluyendo lo relacionado con el COVID
Conocimiento Abierto, transparencia

Tabla XII. Temas predominantes en EPI

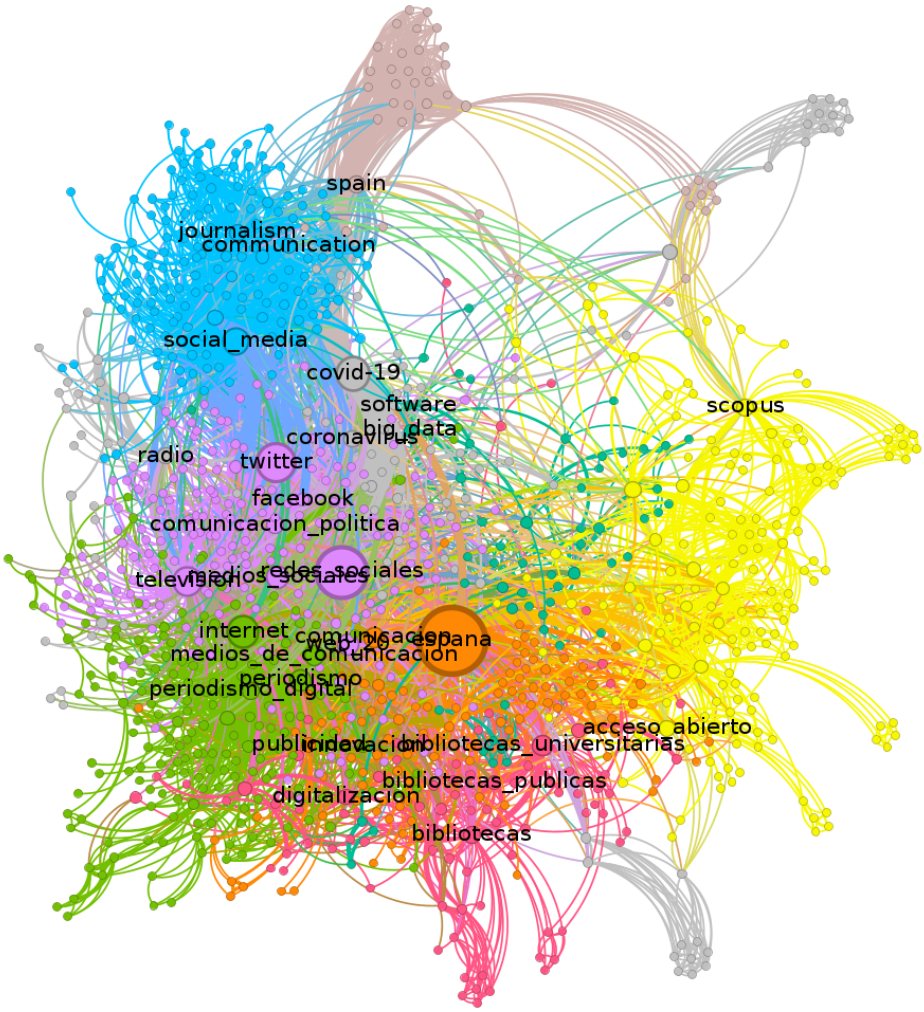

Figura 7. Visualización de la red de palabras clave

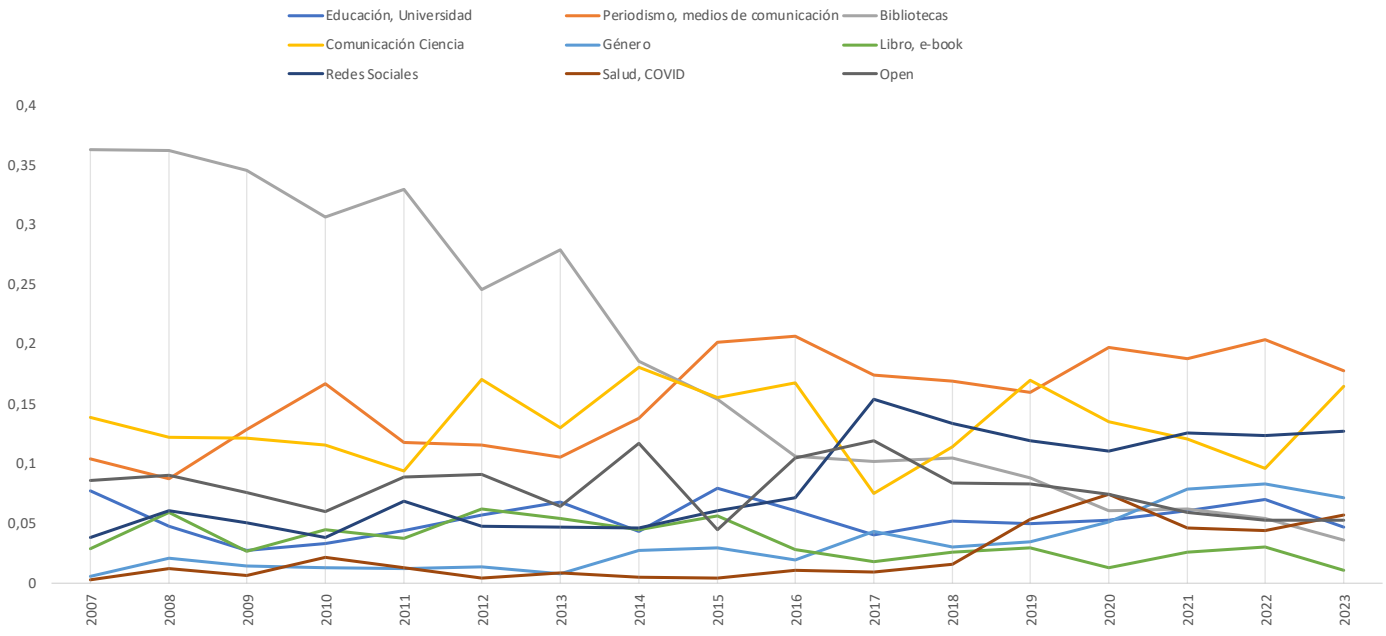

Figura 8. Evolución temporal de los temas

La Figura 9 muestra la importancia o intensidad de cada uno de los nueve temas, mientras que la Figura 8 (en la página anterior) muestra su evolución.

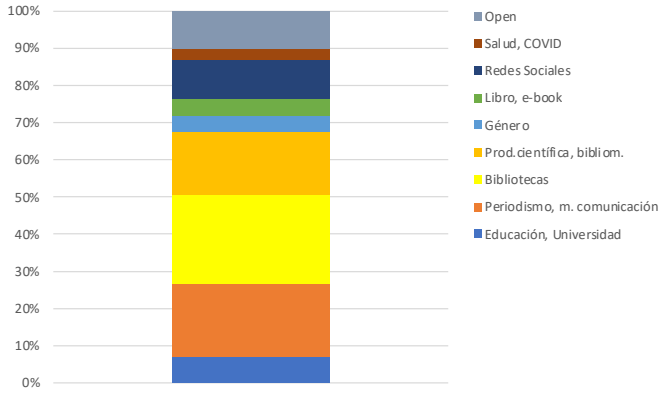

Figura 9. Importancia de los temas principales

Los bloques temáticos, en conjunto, están constituidos por lo relacionado con el periodismo y la comunicación; las bibliotecas y asuntos relacionados; y la bibliometría. Los temas relacionados con las bibliotecas muestran una línea claramente descendente, de manera bastante más suave en los años más recientes. Mientras, lo relacionado con periodismo y medios de comunicación parece evolucionar de manera opuesta. La bibliometría, con picos puntuales, parece tener un nicho bien definido que se mantiene durante todo el período. En los demás temas, menos representados, lo relacionado con las redes sociales muestra un crecimiento notable a partir de 2017; los estudios de género, o relacionados, muy minoritarios hasta 2013, experimentan a partir de entonces un crecimiento considerable.

Lo relacionado con la salud tiene también un incremento notable en los últimos años, relacionado directamente con la pandemia; recordemos también que EPI publica un número especial dedicado a ella. Lo relacionado con el libro y la edición mantiene una línea estable a lo largo del tiempo, al igual que los temas relacionados con la educación y la formación. El conocimiento abierto, en general, parece perder, en los años más recientes, un poco de interés, aunque de forma leve.

5. Conclusiones

La aplicación de las técnicas de análisis de redes sociales permite caracterizar los contenidos de las publicaciones científicas, facilitando la extracción de información, que, de otra manera, dado el volumen de datos existentes, pasaría desapercibida.

A través de las medidas de red de los autores, tenemos unos niveles de conexión extremadamente bajos (densidad), indicando que los niveles de interacción entre autores son relativamente bajos, lo normal son publicaciones entre los mismos grupos de autores. A su vez el alto valor del diámetro es consistente con esta idea.

El promedio de autores ha aumentado en los últimos años, fruto de las colaboraciones más extensas y favorecidas por los equipos multidisciplinares. Las comunidades de autores, nos permiten ver las colaboraciones más fuertes entre los mismos.

Las instituciones a las que pertenecen los autores, si bien no están disponibles en todos los casos, y con las particularidades destacadas, ofrecen un modelo de red en el que podemos ver las instituciones más relevantes, destacando la Universidad Complutense, la Universidad de Granada y la Universidad Pompeu Fabra. La Figura 6 (en la página 21) refleja las interacciones entre dichas instituciones.

Del estudio de las palabras clave y la aplicación de técnicas de detección de topics, hemos obtenido los temas primordiales de la revista. Destacan los temas de bibliotecas, comunicación, redes, producción científica y a pesar de estar limitado en el tiempo el tema sobre el COVID ha dado un impulso a los temas relacionados con información y salud.

Este esquema de trabajo, puede aplicarse a cualquier publicación científica y con la obtención de datos normalizados facilitar la comparación entre publicaciones o analizar una misma publicación a lo largo del tiempo.

Referencias

Adamic, L. A. (2000). Zipf, power-laws, and pareto-a ranking tutorial. Xerox Palo Alto Research Center, Palo Alto, CA, http://ginger. hpl. hp. com/shl/papers/ranking/ranking. html.

Allan, J. (2002). Introduction to topic detection and tracking. // Topic detection and tracking. Springer. 1-16.

Alonso Berrocal, J. L. (2022). Análisis de coautoría y detección de topics en la revista Fonseca Journal of Communication (2010-2020). // Gutiérrez San Miguel, M. B. (Ed.). La transferencia del conocimiento en las revistas científicas: Estudio de caso Fonseca Journal of Communication. Tirant Humanidades. 17-34.

Alperin, J. P.; Willinsky, J.; Owen, B.; MacGregor, J.; Smecher, A.; Stranack, K. (2019). The Public Knowledge Project. // Connecting the Knowledge Commons: From Projects to Sustainable Infrastructure: The 22nd International Conference on Electronic Publishing–Revised Selected Papers, 151.

Beaver, D. Deb. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. // Scientometrics. 52:3, Article 3. https://doi.org/10.1023/A:1014254214337

Benhardus, J.; Kalita, J. (2013). Streaming trend detection in twitter. // International Journal of Web Based Communities. 9:1, Article 1.

Blei, D. M. (2012). Probabilistic topic models. // Communications of the ACM, 55:4, 77-84.

Cheng, F.-F.; Huang, Y.-W.; Tsaih, D.-C.; Wu, C.-S. (2019). Trend analysis of co-authorship network. // Library Hi Tech. 37:1.

Cronin, B. (2004). Bowling alone together: Academic writing as distributed cognition. // Journal of the American Society for Information Science and Technology. 55:6, Article 6. https://doi.org/10.1002/asi.10406

Ferran-Ferrer, N.; Guallar, J.; Abadal, E.; Server, A. (2017). Research methods and techniques in Spanish library and information science journals (2012-2014). // Information research. 22:1. http://InformationR.net/ir/22-1/paper741.html

Figuerola, C. G.; García Marco, F. J.; Pinto, M. (2017). Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. // Scientometrics. 112, 1507-1535.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. // Sociometry. 40:1, 35-41.

García-Marco, F.-J.; Figuerola, C. G.; Pinto, M. (2020). Análisis de la evolución temática de la investigación sobre Información y Documentación en español en la base de datos LISA mediante modelado temático (1978-2019). // Profesional de la información. 29:4, https://doi.org/10.3145/epi.2020.jul.27.

Gleich, D. F. (2015). PageRank beyond the web. // Sian review. 57:3, 321-363.

Guallar, J.; Ferran-Ferrer, N.; Abadal, E.; Server, A. (2017). Revistas científicas españolas de información y documentación: Análisis temático y metodológico. // Profesional de la información. 26:5, 947-960. https://doi.org/10.3145/epi.2017.sep.16

Guallar, J.; López-Robles, J.-R.; Abadal, E.; Gamboa-Rosales, N.-K.; Cobo, M.-J. (2020). Revistas españolas de Documentación en Web of Science: Análisis bibliométrico y evolución temática de 2015 a 2019. // Profesional de la Información. 29:6. http://dx.doi.org/10.3145/epi.2020.nov.06

Guerrero-Castillo, P.; Nuño-Moral, M.-V.; Guerrero-Bote, V. P.; De-Moya-Anegón, F. (2023). New map of the research published in Profesional de la Información (2006-2023). // El Profesional de la información. 32:7, e320708. https://doi.org/10.3145/epi.2023.dic.08

Haak, L. L.; Meadows, A.; Brown, J. (2018). Using ORCID, DOI, and other open identifiers in research evaluation. // Frontiers in Research Metrics and Analytics. 3, 28.

Henriksen, D. (2016). The rise in co-authorship in the social sciences (1980–2013). // Scientometrics. 107:2, Article 2.

Kim, J.; Diesner, J. (2015). The effect of data pre-processing on understanding the evolution of collaboration networks. // Journal of Informetrics. 9:1, Article 1.

Kong, X.; Mao, M.; Jiang, H.; Yu, S.; Wan, L. (2019). How does collaboration affect researchers' positions in co-authorship networks? // Journal of Informetrics, 13:3, Article 3.

Leydesdorff, L.; Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. // Journal of the American Society for Information Science and Technology. 57:12, Article 12. https://doi.org/10.1002/asi.20335

Leydesdorff, L.; Wagner, C. S. (2008). International collaboration in science and the formation of a core group. // Journal of Informetrics. 2:4, Article 4. https://doi.org/10.1016/j.joi.2008.07.003

López-Robles, J.-R.; Guallar, J.; Otegi-Olaso, J.-R.; Gamboa-Rosales, N.-K. (2019). El profesional de la información (EPI): Bibliometric and thematic analysis (2006-2017). // El profesional de la información, 28:4, e280417. https://doi.org/10.3145/epi.2019.jul.17

Lozano, S.; Calzada-Infante, L.; Adenso-Díaz, B.; García, S. (2019). Complex network analysis of keywords co-occurrence in the recent efficiency analysis literature. // Scientometrics. 120, 609-629.

McCallum, A. K. (2002). Mallet: A machine learning for languagetoolkit. http://mallet. cs. umass. edu.

Medrano, J. F. (2020). Enfoque híbrido para la correcta identificación de autores en bases de datos bibliográficas de libre acceso: El caso de Google Scholar. // XXI Simposio Argentino de Inteligencia Artificial (ASAI 2020)-JAIIO 49 (Modalidad virtual).

Molina, J. L. (2001). El análisis de redes sociales: Una introducción. Ediciones Bellaterra.

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. // Proceedings of the national academy of sciences. 101(suppl 1), Article suppl 1.

Olmeda-Gómez, C.; Ovalle-Perandones, M.-A.; Perianes-Rodríguez, A. (2017). Co-word analysis and thematic landscapes in Spanish information science literature, 1985–2014. // Scientometrics. 113:1, 195-217. https://doi.org/10.1007/s11192-017-2486-8

Ossenblok, T. L. B.; Verleysen, F. T.; Engels, T. C. E. (2014). Coauthorship of journal articles and book chapters in the social sciences and humanities (2000-2010): Coauthorship of Journal Articles and Book Chapters in the Social Sciences and Humanities (2000-2010). // Journal of the Association for Information Science and Technology. 65:5, Article 5. https://doi.org/10.1002/asi.23015

Perianes-Rodríguez, A.; Olmeda-Gómez, C.; Moya-Anegón, F. (2010). Detecting, identifying and visualizing research groups in co-authorship networks. // Scientometrics. 82:2, Article 2. https://doi.org/10.1007/s11192-009-0040-z

Peset, F.; Garzón-Farinós, F.; González, L.-M.; García-Massó, X.; Ferrer-Sapena, A.; Toca-Herrera, J. L.; Sánchez-Pérez, E. A. (2020). Survival analysis of author keywords: An application to the library and information sciences area. // Journal of the Association for Information Science and Technology. 71:4, 462-473.

Plantié, M.; Crampes, M. (2012). Survey on social community detection. // Social media retrieval. 65-85. Springer.

Stevenson, W. B.; Greenberg, D. (2000). Agency and Social Networks: Strategies of Action in a Social Structure of Position, Opposition, and Opportunity. // Administrative Science Quarterly. 45:4, Article 4. https://doi.org/10.2307/2667015

Traag, V. A.; Waltman, L.; Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. // Scientific Reports. 9:1, 5233. https://doi.org/10.1038/s41598-019-41695-z

Underwood, T. (2012). Topic modeling made just simple enough. // The Stone and the Shell, 7.

Wasserman, S.; Faust, K. (2013). Análisis de redes sociales: métodos y aplicaciones (Vol. 10). Madrid: CIS-Centro de Investigaciones Sociológicas.

Weibel, S. L.; Koch, T. (2000). The Dublin core metadata initiative. // D-lib magazine. 6:12, 1082-9873.

Wuchty, S.; Jones, B. F.; Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. // Science. 316(5827), Article 5827. https://doi.org/10.1126/science.1136099

Enviado: 2024-03-31. Segunda versión: 2024-05-13.
Aceptado: 2024-05-17.

Inteligencia artificial para el acceso a documentación jurídica y la realización de actividades judiciales

Artificial intelligence and access to legal documentation and carrying out judicial activities

Fernando GALINDO AYUDA

Universidad de Zaragoza, Pedro Cerbuna 12, 50009 Zaragoza, España, cfa@unizar.es.
Miembro de la Real Academia de Jurisprudencia y Legislación de España

Resumen

Este artículo tiene por objeto expresar la relevancia que en las actividades jurídicas, realizadas por profesionales del Derecho, tiene el acceso a documentación jurídica. Ello se muestra tomando en consideración experiencias, normativa y propuestas doctrinales, que se están manifestando y desarrollando, especialmente, en Estados Unidos y España, con respecto a algunas de las posibilidades que ofrece a dicho acceso el uso de la inteligencia artificial, una vez hayan sido desarrolladas al efecto adecuadas aplicaciones o programas.

Palabras clave: Acceso a textos jurídicos. Actividades jurídicas. Aplicaciones de inteligencia artificial. ChatGPT. Reglamento europeo de inteligencia artificial.

Abstract

This article aims to express the relevance that the access to legal documentation has in legal activities, carried out by legal professionals. It is shown by taking into consideration experiences, regulations and doctrinal proposals, which are being expressed and developed, especially in the United States and Spain. These reflections are made with respect to some of the possibilities that the use of artificial intelligence offers to the access to legal documentation, once it exists suitable applications or programs that they have been developed for this purpose.

Keywords: Access to legal texts. Legal activities. Artificial intelligence applications. ChatGPT. European regulation on artificial intelligence.

1. Introducción

El presente trabajo tiene por objeto fijarse en la relevancia que tienen en las sociedades democráticas la organización social denominada Estado de Derecho, y el papel que toma en su puesta en práctica la justificación de las actividades que desarrollan los gobernantes, los profesionales del Derecho y los mismos ciudadanos, mediante las regulaciones que se dan las sociedades democráticas a través de los canales o procedimientos establecidos por el mismo Estado de Derecho.

Esta justificación sucede, fundamentalmente, gracias a lo que prescribe el texto de las leyes. Aquí consideramos incluido en el texto de las leyes al de las mismas leyes, al de la Constitución, y al de su concreción hecha por: las resoluciones judiciales, las regulaciones y decisiones administrativas que realizan los miembros del poder ejecutivo, o los mismos textos expresados en forma de contratos y convenios, que formulan el contenido de las voluntades de los ciudadanos acordadas con otros ciudadanos.

Es conveniente indicar aquí que las anteriores afirmaciones, o lo que es lo mismo la fundamentación en el texto de la Ley del Derecho, están ligadas, culturalmente, a la concepción sobre el sistema jurídico denominada “Derecho Civil”, “Derecho continental” (o Derecho de tradición romano-germánica) —el caso de España—; a diferencia de lo que sucede en los países anglosajones en los que la concepción cultural del Derecho es la de “Common law”, expresión que es usada para referirse al sistema jurídico del mismo nombre, cuyo ejercicio está basado, primordialmente, en las decisiones adoptadas por los tribunales.

Sobre los orígenes y, a la vez, las características del “Common law” y su relación con la sociedad en la que se constituye, se habla detalladamente en el libro de Russell Sandberg. Esto es, según el autor expone, porque la historia refleja las relaciones existentes entre el cambio jurídico y las influencias externas sociales y políticas (Sandberg, 2023, p. 10). Ello contrasta con los sistemas de “Derecho continental” donde la principal fuente de Derecho es la Ley. En este trabajo, como hemos dicho, nos fijamos preferentemente en la perspectiva propia de los sistemas de “Derecho continental” o “Derecho Civil”.

Desde el expresado punto de partida es importante poner de manifiesto la relevancia que alcanza, en la vida de las sociedades democráticas,

el conocimiento de los textos jurídicos por los agentes jurídicos, que se encargan de poner en acción los textos de las leyes ante cualquier eventualidad o ejercicio de derechos que comporte la exigencia de su cumplimiento o la aparición de concretas incidencias o problemas. Todo esto cabe expresarlo, resumidamente, como la necesidad de acceder a textos jurídicos.

Es preciso mencionar que aquí no nos referimos al acceso a textos jurídicos por los ciudadanos. Este es tema del que se ocupan libros clásicos como los titulados "El Derecho en casa" o similares, que a lo largo de los siglos (ya desde el siglo XVI existen ejemplares con esta denominación: uno de los más antiguos es obra de Francisco de Vitoria, fue publicado impreso en 1532) se editaron a efectos de que los ciudadanos pudieran conocer el contenido y características del Derecho en términos próximos a ellos. Un ejemplo reciente de este tipo de aproximación lo constituye el libro de Alfred Font y José Luis Pérez, que atiende a las necesidades de "un lector no especializado" (Font, Pérez, 2009, p.11).

El acceso a textos jurídicos ha tenido lugar en las sociedades democráticas en forma congruente a lo que requería el formato en el que dichos textos estaban recogidos. El desarrollo de la imprenta desde el Renacimiento permitió que esos textos estuvieran impresos en libros o papeles en forma de códigos aprobados, progresivamente, por las autoridades democráticas, siendo por ello accesibles directamente por quienes supieran leer o estudiar su contenido, sin contar con otros intermediarios que quienes los hicieran, o intervinieran en su publicación. En la actualidad los textos impresos son accesibles y legibles predominantemente por la publicación de los mismos que se produce gracias a la existencia de Internet y las tecnologías de la información y la comunicación.

Se habla de "actividades jurídicas" porque el acceso a textos jurídicos es una actividad jurídica a la que cabe diferenciar de otras como son la interpretación y la aplicación del Derecho, realizadas también por juristas. La realización de estas actividades también es auxiliada hoy por las tecnologías de la información y la comunicación.

La consideración del auxilio de las tecnologías a estas otras actividades (interpretación y aplicación) no es objeto de este trabajo con una única excepción: aquí realizamos una atención, resumida, a varias posiciones mantenidas a favor y en contra de la realización de actividades jurídicas por los jueces con auxilio de dichas herramientas tecnológicas y usando las posibilidades que ofrece la denominada inteligencia artificial (de ahora en adelante IA). Esto lo hacemos en el apartado 6 a efectos de comprender con ello la regulación hecha al respecto por el Reglamento europeo sobre inteligencia artificial, en el que se establece una diferente atención a las actividades de aplicación e interpretación del Derecho y su uso de la IA, que la que se hace al ejercicio de la actividad del acceso a textos jurídicos mediante el uso de esta técnica, regulación que concretamos en el apartado 5 al exponer el contenido fundamental del Reglamento europeo de inteligencia artificial en lo referido al uso de la IA en la realización de las actividades jurídicas .

Como métodos adecuados para discurrir sobre lo hasta aquí expresado tomamos como referencia los siguientes. Tras efectuar una consideración introductoria sobre la necesidad estipulada en el Estado de Derecho de motivar las resoluciones jurídicas, realizamos los siguientes estudios. Primeramente: 1) un estudio sobre el contenido básico de los correspondientes textos promulgados sobre el acceso a textos jurídicos realizado, especialmente, con auxilio de las tecnologías de la información y comunicación, y 2) un estudio del contenido de algunas recientes opiniones doctrinales expresadas sobre las características y calidad del acceso a textos jurídicos. En segundo lugar, tenemos en cuenta lo que aportan datos concretos referidos al estado de implantación y desarrollo de varios recursos técnicos, de carácter público o privado, que auxilian al acceso a textos jurídicos utilizando la IA.

Las aportaciones quedan hechas mediante el contenido expresado en los diferentes apartados de este trabajo titulados de esta forma: la motivación de las actividades jurídicas en el Estado de Derecho (apartado 2); la necesidad de acceder a textos jurídicos (apartado 3); las posibles virtualidades de la inteligencia artificial para dar acceso a textos jurídicos (apartado 4); el reglamento europeo de inteligencia artificial en lo relativo al uso de la IA en el acceso a textos jurídicos y en la realización de actividades judiciales (apartado 5); el uso de la inteligencia artificial por las actividades judiciales (apartado 6).

El trabajo termina con la conclusión, (apartado 7), que concreta una respuesta a la siguiente pregunta: ¿el uso de las tecnologías de la información y la comunicación en el acceso a textos jurídicos, a juristas interesados, cambia sus exigencias si se hace realidad el desarrollo de aplicaciones o programas de ordenador denominados de IA para dicha actividad?

2. La motivación de las actividades jurídicas en el Estado de Derecho

Desde el Renacimiento se fue elaborando la idea de que la acción del poder político debe fundamentarse en la voluntad de los ciudadanos que

integran la sociedad en la que gobierna dicho poder. Con ello se irían dejando de lado, al contrario de lo que sucedía con anterioridad, los fundamentos teológicos del poder (Habermas, 2023, p. 703):

> La teoría moderna del conocimiento elabora la independización de la astronomía y la física […] mientras que el derecho racional reacciona a la doble confesionalización de la fe y a la violencia de las guerras de religión con proyectos de constitución que tenían que *constitucionalizar* el ejercicio de la soberanía política independientemente del derecho natural divino o cosmológico.

La puesta en realidad de estas ideas no fue lograda hasta que en el siglo XVIII no se produjeron las revoluciones americana y francesa, cuando los ciudadanos crearon el nuevo poder político que se estaba erigiendo en América frente a la metrópoli, o, en Francia, los representantes de los ciudadanos se constituyeron en Asamblea y República frente al monarca, cuyo uso del poder o la fuerza estaba legitimado hasta entonces por la tradición y, al final, por el mismo Dios.

Los efectos de estas revoluciones se harán realidad en otros países desde el siglo XIX hasta la actualidad, siendo aceptada en numerosos lugares la fundamentación racional / contractual del poder político y de las relaciones establecidas entre ciudadanos. Otro principio también aceptado en la actualidad es el de la organización del ejercicio del poder político mediante la separación de poderes: el legislativo, el ejecutivo y el judicial.

Efecto concreto de estas ideas, especialmente en el ámbito cultural de los países de "Derecho Civil", es el reconocimiento de que las decisiones judiciales han de estar motivadas o fundamentadas por el texto de las leyes. Así la Constitución española de 1978 dice en su art. 120.3: "Las sentencias serán siempre motivadas y se pronunciarán en audiencia pública".

O, lo que es lo mismo, no serán "fruto de un mero voluntarismo selectivo o de la pura arbitrariedad". Así se recoge en la Sentencia del Tribunal Supremo, Sala tercera, de 12 de febrero de 2008:

> El derecho a la tutela judicial efectiva no exige que la resolución judicial ofrezca una exhaustiva descripción del proceso intelectual llevado a cabo por el juzgador para resolver, ni una pormenorizada respuesta a todas las alegaciones de las partes, ni siquiera la corrección jurídica interna de la fundamentación empleada, bastando con que la argumentación vertida exteriorice el motivo de la decisión, la ratio decidendi, en orden a un eventual control jurisdiccional, pues se cumple la exigencia constitucional cuando la resolución no es fruto de un mero voluntarismo selectivo o de la pura arbitrariedad.

Lo mismo sucede con los actos administrativos. En efecto: en el art. 35 de la Ley 39/2015, de 1 de octubre, del Procedimiento Administrativo Común de las Administraciones Públicas se establece detalladamente la obligación del órgano administrativo que adopta la decisión de incluir en ella una exposición sucinta de los hechos y fundamentos jurídicos en que se basa.

Igualmente es así con los contratos o acuerdos realizados entre ciudadanos, que han de "establecer los pactos, cláusulas y condiciones que tengan por conveniente, siempre que no sean contrarios a las leyes, a la moral, ni al orden público", tal y como expresa el art. 1.255 del Código Civil vigente, promulgado en 1889.

3. La necesidad de acceder a textos jurídicos y el uso de medios telemáticos en la Administración de Justicia

La necesidad de acceder a textos jurídicos para su interpretación, aplicación o elaboración de otros textos, procede de las revoluciones, los principios y las regulaciones resumidas en el anterior apartado. Ha quedado satisfecha a través del funcionamiento de las instituciones y los procedimientos adoptados para su puesta en práctica, contemplando las posibilidades ofrecidas por las técnicas, la imprenta, especialmente. Potencialidades que están, desde hace ya cierto tiempo, incrementadas por lo que permite el uso de las tecnologías de la información y la comunicación, que se han mostrado especialmente adecuadas para almacenar y acceder a textos jurídicos.

En España con respecto al uso de estas tecnologías, tras la aprobación de la Constitución y su puesta en acción a través de la promulgación de las normas precisas para hacer realidad los principios señalados, se satisfizo dicha necesidad en el ámbito de las actividades jurídicas atendiendo especialmente a los procesos y procedimientos que tienen lugar en el ámbito judicial. Así en la Presentación del libro titulado *Gestión automatizada en el ámbito de la Justicia* (CREI, 1983, p. 7):

> Desde los inicios de la informática moderna, alrededor de 1950 […] esta ciencia ha ido invadiendo todas las áreas del tradicional saber humano […] Haciendo un sucinto inventario de las influencias de una ciencia [la informática] sobre la otra [el derecho] y de sus aplicaciones nos vemos obligados a hablar ya de diversas ramas: Informática Registral y Documental Jurídicas, en las que encontramos aplicaciones de la informática para la clasificación y tratamiento de grandes bancos de datos jurídicos, como los de Registros de Penados y Rebeldes, los Registros de últimas Voluntades […], o bien como pueden ser archivos o bases de datos que contengan la Legislación, la Jurisprudencia y la Doctrina Jurídica, permitiendo acceder a ellas en brevísimos espacios de tiempo por medio de sencillos parámetros que nos conducen a la información deseada.

En el libro señalado, publicado a comienzo de los años ochenta del pasado siglo, se concretaban las posibilidades establecidas en la presentación del mismo mediante la recopilación de diferentes trabajos y propuestas, elaboradas por juristas: jueces y, también, funcionarios responsables de dotar a la Administración de Justicia (estatal y autonómica) de la infraestructura precisa para su funcionamiento.

Hoy cabe decir que estas posibilidades han pasado a ser una realidad regulada detalladamente (1). Disposiciones legales y experiencias que han sido recogidas en una norma que, a la vez, ha dictado pautas para su inmediata actuación, a la vez que su futuro. La norma a la que nos referimos es el Real Decreto-ley 6/2023, de 19 de diciembre, por el que se aprueban medidas urgentes para la ejecución del Plan de Recuperación, Transformación y Resiliencia en materia de servicio público de justicia, función pública, régimen local y mecenazgo.

Así, congruentemente con lo dicho, en el Preámbulo de este Real Decreto-ley (apartado II), se reconoce:

> […] que la celebración de vistas y actos procesales mediante presencia telemática, son hoy día parte de la actividad cotidiana del servicio público de Justicia.

Asimismo, se concreta que el objeto principal de la regulación es establecer:

> […] la obligación de las administraciones competentes en materia de Justicia de garantizar la prestación del servicio público de Justicia por medios digitales, equivalentes, de calidad y que aseguren en todo el territorio del Estado una serie de servicios, entre los que se encuentran, como mínimo, (i) la iteneración de expedientes electrónicos y la transmisión de documentos electrónicos entre cualesquiera órganos judiciales o fiscales, (ii) la interoperabilidad de datos entre cualesquiera órganos judiciales o fiscales, (iii) el acceso a los servicios, procedimientos e informaciones de la Administración de Justicia que afecten a la ciudadanía, y (iv) la identificación y firma de los intervinientes en actuaciones y servicios no presenciales.

El párrafo finaliza diciendo:

> El texto normativo se erige como un instrumento para promover y facilitar la intervención telemática de los ciudadanos en las actuaciones judiciales, simplificándose la relación con la Administración de Justicia.

Mecanismos y actividades que, se sobreentiende en todo caso, han de estar motivados o fundamentados por lo que señalan el texto de las leyes: los textos jurídicos que la Constitución y la jurisprudencia del Tribunal Supremo, como antes hemos indicado.

4. Las posibles virtualidades de la inteligencia artificial para dar acceso a textos jurídicos

La norma de 2023 es coherente con el hecho de que desde el comienzo de la generalización del uso de las tecnologías de la información y la comunicación en el ámbito jurídico (desde los años ochenta del siglo XX), la satisfacción de la necesidad de acceder a textos jurídicos se ha producido mediante la construcción de sistemas o programas dedicados al almacenamiento y acceso a dichos textos, gracias a la aportación, realizada por las instituciones generadoras de dichos textos, de bases de datos o programas.

Bases de datos que contienen: la Constitución; las leyes; la jurisprudencia o las decisiones de tribunales o jueces; las decisiones del Tribunal Constitucional; las resoluciones y normas creadas por autoridades administrativas; los mismos formularios o modelos de documentos precisos para poner en práctica los procedimientos de solución de conflictos o la elaboración de normas; los modelos de contratos que se acostumbra a acordar socialmente, e, incluso, los textos de libros y revistas que contienen las reflexiones establecidas al respecto por los estudiosos de los textos jurídicos, que proponen interpretaciones de los mismos integrando las propuestas de la denominada dogmática o ciencia del Derecho.

Es conveniente resumir aquí algunas de las características de estas iniciativas a efectos de poder diferenciarlas de otras que están apareciendo como posibilidades a través del desarrollo de programas de inteligencia artificial (IA). A estos efectos en el presente apartado nos referimos a: 1) El acceso a textos jurídicos mediante la utilización de bases de datos, 2) La caracterización de los programas de inteligencia artificial, y 3) Las propuestas de acceso a textos jurídicos mediante aplicaciones de inteligencia artificial.

4.1. El acceso a textos jurídicos mediante la utilización de bases de datos

No es este el momento adecuado para hacer una historia pormenorizada de la aparición del uso de las tecnologías de la información y la comunicación en el acceso a textos jurídicos. Aquí vamos a centrarnos en hacer un breve resumen de lo que permiten al respecto algunas iniciativas de carácter público y privado que en España se ocupan de proporcionar dicho acceso.

En el ámbito público se destacan las bases de datos que procuran acceso a los textos que crean o generan los organismos oficiales. En lo que se refiere a las leyes aprobadas por el Parlamento, las

Cortes, y las normas promulgadas por las Administraciones se destacan las que contienen el Boletín Oficial del Estado, y los distintos órganos de publicación de las normas aprobadas por cada una de las Comunidades Autónomas. Las sentencias, autos y declaraciones del Tribunal Constitucional son publicadas por el mismo Tribunal. Todas ellas son accesibles y publicadas en Internet.

Siguiendo en el ámbito público. Lo referido a las sentencias dictadas por el Tribunal Supremo, los Tribunales Superiores de Justicia y las Audiencias Provinciales, se publican también en Internet. Constituyen bases de datos, cuya responsabilidad de creación, anonimización (de la información personal —los nombres y apellidos— referida a las partes, ciudadanos, contenida en las sentencias, tal y como establecen las normas de protección de datos personales) y acceso recae en el Centro de Documentación Jurídica (CENDOJ) que es parte integrante del Consejo General del Poder Judicial, órgano de gobierno supremo de los jueces.

En el ámbito privado son varias las empresas que, atendiendo a diferentes perspectivas, aproximaciones y técnicas, proporcionan a sus clientes sistemas de acceso a documentación que almacenan y acceden a lo siguiente: los textos de la legislación y las normas aprobadas por las Administraciones públicas, las resoluciones del Tribunal Constitucional, las sentencias que integran la base de datos de jurisprudencia constituida y proporcionada por el CENDOJ, los convenios colectivos aprobados por trabajadores y empresas, los formularios de documentos que pueden formar parte de procedimientos a plantear (con respecto a casos concretos) ante autoridades públicas (Administraciones o Poder Judicial), también los modelos de contratos utilizados habitualmente por ciudadanos que deciden realizar acuerdos o convenios con otras personas. Dichas bases de datos, igualmente, dan acceso a comentarios doctrinales publicados en revistas o libros.

Estas bases de datos son la versión en formato electrónico de los textos que, en buena parte de las ocasiones, son también publicados en formato papel por las mismas instituciones públicas y empresas. La mayor diferencia entre los textos impresos en papel y los textos digitales o electrónicos reside en que los publicados en este último formato son accedidos mediante lenguajes de interrogación, aplicaciones o programas, denominados sistemas de búsqueda de documentación, que permiten acceder al contenido y forma de los textos jurídicos con mayor agilidad y virtualidad que los que permite la consulta de los textos impresos y publicados únicamente en formato papel.

El acceso se produce contando siempre con la limitación, a la vez que ventaja, propia de los sistemas electrónicos, de que las bases de datos organizan tanto la información almacenada como los sistemas ocupados de su recuperación y acceso, con base a conceptos jurídicos. Esto es: el lenguaje de juristas que está moldeado o formado, básicamente, por los datos propios de cada clase de documento jurídico prescritos por las normas, e incluso establecidos por la dogmática o la ciencia del Derecho. Es por ello claro que las leyes cuenten con un lenguaje "normal" y un lenguaje "técnico-jurídico". Así dice Gregorio Robles (Robles, 2021, p. 394) que atender al lenguaje "técnico-jurídico" es preciso porque está integrado por:

> […] los términos y modos expresivos propios de los juristas y que han aprendido en su paso por la Facultad de Derecho y en lecturas y estudios posteriores. Es evidente que esos términos y modos expresivos forman parte del lenguaje de los juristas y puede suceder que sea conveniente e incluso necesario introducir esos términos en las leyes.

A lo anterior no le impide reconocer que en la mayor parte de las ocasiones también es posible acceder a los textos jurídicos utilizando texto libre.

El problema común que tiene este sistema de acceso está centrado en que sus respuestas (en definitiva, el conjunto de documentos jurídicos que contienen las expresiones de búsqueda) no son satisfactorias por requerir del interesado que realiza la consulta un trabajo ímprobo dada la gran cantidad de respuestas que puede producir el sistema de acceso por el uso del texto conceptual o, mucho más, el texto libre como recurso de acceso.

Esto hace que los sistemas no permitan de hecho acceder a la documentación atendiendo únicamente a las particularidades del lenguaje común utilizado por ciudadanos o personas no expertas en Derecho. Dificultades que, como hemos dicho, también tienen los juristas. En este caso la razón de las mismas reside en que aun consultando la información conociendo y usando el lenguaje jurídico, es usual que las respuestas que ofrezca el sistema sean numerosas, y muchas de ellas no pertinentes.

Lo anterior implica que una vez accedidos los textos han de ser trabajados / estudiados por los juristas, con el fin de poder ser utilizados adecuadamente en apoyo de los argumentos que deberá contener su propuesta de resolución de los casos, asuntos o conflictos concretos, por los que hayan hecho la consulta.

4.2. La caracterización de los programas de inteligencia artificial

4.2.1. Generalidades

El problema que acaba de señalarse depende del mecanismo de funcionamiento de los programas informáticos que son los sistemas de recuperación y búsqueda de información en las bases de datos jurídicas indicadas.

Estos programas, guiados por los algoritmos de funcionamiento de los mismos, se encargan de realizar cálculos con el objetivo de proponer cuáles de los numerosos textos almacenados en el sistema coinciden con los propuestos por quien hace la consulta. Por ello los textos de la respuesta pueden, o no, ser adecuados para ayudar a resolver el problema concreto del que se ocupa en su trabajo profesional el jurista que pregunta a la base de datos mediante el programa de búsqueda.

Estos cálculos se satisfacen, cumpliendo con el algoritmo correspondiente, con datos que conceptualizan / expresan / resumen la información que se almacena en las que se denominan bases de datos, o simplemente datos, que en la actualidad conforman en buena medida todo lo que hoy constituye Internet.

Tras el hallazgo de los datos / textos viene, por tanto, el trabajo jurídico de interpretarlos para ver cuáles de ellos suministran la fundamentación de la posición por la que el jurista haya realizado la búsqueda de información.

Unas palabras sobre qué es el algoritmo. Donald Knuth, considerado el padre de la programación informática moderna, dice que es el conjunto de instrucciones que se deben seguir para resolver un problema o completar una tarea. Más en concreto, Knuth define que el algoritmo es (Knuth,1997, p. 4):

> […] simplemente un conjunto finito de reglas que da una secuencia de operaciones para resolver un tipo específico de problema.

Los algoritmos se usan en muchos campos diferentes, así como en la ciencia, la ingeniería, la informática, los negocios… También, como hemos señalado, en los programas que se utilizan como instrumento auxiliar en las actividades profesionales de los juristas con el fin de encontrar y proporcionar aquellos textos jurídicos que puedan servir para motivar sus trabajos o argumentos.

¿Qué hacen, además, las herramientas de inteligencia artificial (IA)? De acuerdo con la literatura, como se describe, por ejemplo, en Russel y Norvig (2021, p. XI-XVII), estos programas también están destinados al cálculo de datos, posibilitando lo siguiente: el aprendizaje automático, el procesamiento del lenguaje natural, la visión por computador y el funcionamiento de la robótica. Esto es: son herramientas que se utilizan para hacer: programas / construcciones / fórmulas / cálculos / información / números / algoritmos / textos / bases de datos. A ello no obsta reconocer que, en el ámbito jurídico, como en otros ámbitos sociales, se han hecho, desde hace tiempo, aplicaciones de las diferentes herramientas de IA que acaban de indicarse.

4.2.2. ChatGPT

Importa señalar en este momento que, desde fechas recientes, existe una aplicación de IA que está tomando una gran atención e interés por quienes se ocupan de suministrar el acceso a la documentación jurídica, tema que hemos considerado objeto de este trabajo. Es la aplicación denominada “Chat GPT”. A continuación se explica sucintamente su función.

Con la expresión “chat” se expresa la acción de charlar / conversar / argumentar.

Más en concreto: según la Real Academia española chatear es “mantener una conversación mediante chats”. Y chat es “el intercambio de mensajes electrónicos a través de internet que permite establecer una conversación entre dos o más personas”.

El programa ChatGPT ha sido desarrollado desde hace poco tiempo (finales de 2022) en los Estados Unidos.

Es un modelo de IA, diseñado por la empresa estadounidense “OpenAI”, que puede generar respuestas a preguntas o incluso contenido en lenguaje natural como si fuera un ser humano, tal y como expresan los autores que han desarrollado el programa (2).

Está fundamentado en los avances logrados por las herramientas de IA denominadas procesamiento del lenguaje natural (fundamentalmente del lenguaje natural inglés, para ser más exactos, que es el idioma que en mayor medida se usa en Internet y por ello el que más estudiado está), y del aprendizaje automático.

Para conocer la evolución de los trabajos realizados en relación a la inteligencia artificial y el lenguaje es de interés el trabajo de Asunción Gómez Pérez (Gómez-Pérez, 2023, p. 40-83). Una reciente propuesta (Colombo, 2024) discurre sobre la construcción de un modelo de procesamiento del lenguaje natural inglés actual en el ámbito jurídico. Ha de hacerse constar que las normas jurídicas consideradas en la propuesta proceden

de países de common law, y del derecho promulgado por la Unión Europea —normas emanadas del Parlamento y del Consejo europeos—, que es de tipo continental (se puede ver la lista de la normativa considerada en Colombo, 2024, p. 3).

Continuando con las funcionalidades de ChatGPT, ha de decirse que es el programa que adjetiva a las conversaciones o diálogos, Chats, creados por lo que indica la expresión "Generative Pre-trained Transformer" (GPT). En español, literalmente: "Transformador Generativo Pre-entrenado". La idea se contiene también en la siguiente expresión utilizada en español para denominar a la herramienta: "Inteligencia artificial generativa". Términos que muestran la dependencia del programa del desarrollo de las técnicas de IA antes señaladas.

Las denominaciones expresan que el programa ChatGPT es un modelo de lenguaje entrenado con gran cantidad de datos de texto, accesibles en Internet por ser de dominio público o, en algunos casos, por haber obtenido el permiso de sus propietarios (3), para poder realizar conversaciones con los usuarios en una amplia variedad de tareas humanas.

Esto significa, según dicen sus creadores, que su capacidad tecnológica para comprender el contexto y la intención existentes detrás de las preguntas o consultas de los usuarios lo convierten en una herramienta que puede desarrollar conversaciones ("chats") con "robots" o sistemas informáticos y, con ello (por ejemplo: en relación al tema del que nos ocupamos) mejorar la precisión en el uso de los sistemas de búsqueda de información afinando las respuestas que proporcionan los sistemas de acceso a documentación jurídica incrementando con ello su efectividad. Todo lo cual no impide reconocer, como se explica en las condiciones de uso del programa (nos fijamos en las que entraron en vigor el 31 de enero de 2024) al explicar el Contenido que proporciona el sistema, más en concreto: la exactitud de su respuesta o "output", que hay que tener precaución con estas respuestas porque, como se indica al usuario en https://openai.com/policies/terms-of-use:

> El Output puede no ser siempre exacto. No considere que el Output de nuestros Servicios es la única fuente de información veraz o fáctica, ni un sustituto del asesoramiento profesional.

En el próximo subapartado recogemos algunos ejemplos sobre el alcance, que los programas de IA declaran estar logrando en la actualidad, en lo relativo a su uso en el acceso a textos jurídicos.

4.3. Las propuestas de acceso a textos jurídicos mediante aplicaciones de inteligencia artificial

En este lugar nos vamos a referir al alcance y a las posibilidades que tiene la IA en el acceso a textos jurídicos, atendiendo a algunas manifestaciones que se están produciendo al respecto desde hace no mucho tiempo.

Los ejemplos versan sobre lo siguiente: primero, se expresa el objeto de varias experiencias que tienen lugar en el ámbito cultural de "Common law" (Estados Unidos y Reino Unido) (4.3.1); segundo, en España, como ejemplo de un país representante del ámbito del Derecho continental (4.3.2).

4.3.1. Experiencias en el ámbito cultural de Common law

Resumimos aquí el objetivo y la función que se persigue por varias experiencias que se están produciendo en relación al tema que nos ocupa en países de Common law especialmente. Este objetivo da significativa cuenta del alcance de sus logros.

En primer lugar, hacemos mención a lo que declaran experiencias referidas al auxilio de la IA a la revisión de contratos. En segundo lugar, nos referimos a aplicaciones de IA que auxilian a la actuación en procedimientos procesales. En tercer lugar, realizamos algunas consideraciones generales sobre las características y funciones de este tipo de programas recogidas en los dos apartados anteriores.

De la revisión de contratos se ocupa un sistema denominado "autorización de la revisión de contratos". Este sistema aprovecha, mediante la aplicación de la tecnología, desarrollada por la plataforma "Lawgeex" (creada y actualizada en Estados Unidos e Israel) que lo ofrece, las virtualidades de la IA para revisar y marcar documentos legales en función de las políticas predefinidas por quien contrata su utilización. A diferencia de otras soluciones que solo señalan cláusulas inaceptables o faltantes, el programa entiende el contexto contractual, así como la posición de quien encarga el uso del sistema.

El sistema establece, por tanto, líneas rojas en el contrato y negocia con la contraparte, al igual que un abogado experimentado.

De esta forma la IA ayuda a los equipos jurídicos de las firmas que usan el sistema, a automatizar el proceso de revisión del contrato durante la fase previa a la firma, entendiendo el contexto contractual y la posición de quien encarga la tarea. Según resume la plataforma (4):

> Our technology makes redlines to the contract and negotiates with the counterparty – just like an experienced attorney, but with enhanced speed and accuracy.

El Sistema, por tanto, complementa las propuestas que hace un programa tradicional de acceso a documentación jurídica, auxiliando al incremento de la eficiencia del trabajo propio de los juristas que lo utilizan.

Otros sistemas auxilian a la realización de procesos. Este es el caso del programa que ofrece la empresa Luminance (situada en Londres). Esta empresa, fundada por expertos en inteligencia artificial de la Universidad de Cambridge, se constituye como plataforma de IA para abogados.

La plataforma, ofertada con el mismo nombre de la empresa, está basada en un modelo de procesamiento del lenguaje natural legal que la empresa patentó. El sistema lee y construye una comprensión conceptual de documentos jurídicos. A partir de este entendimiento, Luminance mejora y agiliza una amplia gama de tareas, por ejemplo: realizar una revisión inicial de cualquier contrato entrante y marcar automáticamente anomalías contractuales; resaltar áreas de incumplimiento que deben remediarse y etiquetar cláusulas. De esta forma el sistema lleva la IA a cada punto de contacto que un abogado tiene con sus contratos. Como la página web dice sobre el sistema (5):

> Luminance's specialist 'legal-grade' AI is a trusted co-pilot for any legal team.

Como el caso anteriormente mencionado el sistema de IA ayuda más eficazmente que los sistemas tradicionales de acceso a bases de datos jurídicas a la realización de actividades propias de los despachos y oficinas de abogados en relación a su intervención en procesos judiciales.

Ha de tenerse en cuenta para la valoración de estas experiencias, y de su posible alcance, que, sin duda, las propuestas son reales: son ofertadas en Internet por empresas para su utilización directa a los clientes, abogados, interesados, previo el correspondiente pago de los servicios prestados.

Como hemos visto los ejemplos presentados abordan siempre problemas concretos, convenientemente delimitados: especialmente la elaboración de contratos en conformidad con experiencias, casos o precedentes. También auxilian a las actividades propias de oficinas de abogados en relación a la gestión de procesos judiciales.

Ha de considerarse que los sistemas mencionados están caracterizados por aplicar el fuerte desarrollo tecnológico logrado con respecto al procesamiento de lenguaje natural inglés, lenguaje usado mayoritariamente en Internet.

Finalmente, no hay que olvidar que las iniciativas que tienen lugar en el ámbito jurídico señalado siempre tienen presente que el Derecho tiene como referencia cultural el funcionamiento del Derecho de "Common law", de perfiles muy distintos al propio de países, como es el caso de España, en los que su cultura jurídica tiene carácter continental. Recuérdese que la distinción de los sistemas depende del énfasis que el primero pone en la jurisprudencia como precedente, y el segundo en la fundamentación de las actividades jurídicas en el texto de las leyes de referencia en el caso o asunto que las genera.

4.3.2. Situación en España

En este lugar nos fijamos en la atención que se pone en España, mostrando algunos ejemplos, en el ámbito jurídico a las aplicaciones de programas de IA en relación al acceso a textos jurídicos. Tomamos como referencia comparativa los desarrollos de programas de acceso a textos jurídicos elaborados a partir de la utilización de sistemas informáticos tradicionales a cuyo funcionamiento nos hemos referido, sintéticamente, más arriba (4.1).

Distinguimos entre las iniciativas que se ocupan de resolver asuntos generales de acceso a los textos jurídicos mediante el auxilio de la IA, y las que se ocupan de auxiliar con esta herramienta a actividades ligadas a procesos judiciales concretos.

Al final establecemos algunas consideraciones generales sobre lo expresado.

4.3.2.1. Asuntos generales

En lo referido a la resolución de asuntos generales hay que decir lo siguiente.

1) En el ámbito público la novedad más destacada procede de la actuación del CENDOJ, órgano del Consejo General del Poder Judicial que se ocupa de proveer de los textos de sentencias judiciales a los ciudadanos, juristas y empresas de documentación que proporcionan dichas sentencias. La novedad reside en que el CENDOJ, en lo referido a su labor de anonimización de las sentencias, ha desarrollado un programa denominado KENDOJ que utiliza técnicas de IA. KENDOJ significa "Knowledge Extractor for CENDOJ", Por este programa el CENDOJ automatiza dicha anonimización pasando de realizar la acción "manualmente" a hacerlo automáticamente.

Con respecto a este programa en la correspondiente Memoria del Consejo General del Poder Judicial (Memoria, 2022, p. 250) se dice:

> En este año 2021 se ha realizado la implementación de KENDOJ y puesta a disposición de la carrera judicial, Letrados/as de la administración de justicia,

> fiscales y abogados/as del estado. KENDOJ es una herramienta desarrollada con técnicas de inteligencia artificial, de procesamiento de documentos jurídicos en español, fundamentalmente resoluciones judiciales que permite, partiendo de un documento Word, (RTF), estructurarlo, detectar entidades, seudonimizarlo y validar dicha seudonimización, detectar vínculos a jurisprudencia y legislación y clasificarlo en base a voces del Tesauro jurídico Cendoj. Tecnológicamente KENDOJ sigue un enfoque NLP mixto, combinando sistemas de reglas y diccionarios con procesos de "machine learning", y está construido sobre herramientas "open source" de la "Apache Software Foundation", como pueden ser Apache UIMA, Apache Tika o Apache OpenNLP.

2) La mayoría de las empresas que se dedican en España a proporcionar acceso a textos jurídicos a sus clientes realizan también propuestas referidas a su voluntad de usar la IA. Incluso en varios casos la referencia es al uso de la "inteligencia artificial generativa" (ChatGPT), como complementa a la información o datos contenidos en sus programas dedicados a dicho acceso.

A continuación, resumimos algunas de las iniciativas o afirmaciones que hacen al respecto en sus páginas web las empresas o plataformas más significativas en España que se dedican a construir y ofertar programas de acceso a documentación jurídica.

2.1) Hay una propuesta de la empresa Wolters Kluwer (La Ley) que interesa destacar en primer lugar. Es la presentación de los resultados de una encuesta hecha por la empresa en 2023 a abogados sobre varios temas, entre ellos sobre el uso de IA.

Se trata de la encuesta titulada "El abogado del futuro 2023", que la empresa realizó a setecientos profesionales del sector jurídico que trabajan en despachos de abogados y departamentos jurídicos corporativos en Europa (Alemania, Países Bajos, Reino Unido, Bélgica, Francia, Italia, España, Polonia y Hungría) y EE. UU. (6).

En relación a las respuestas de los abogados con relación a IA se obtuvieron las siguientes conclusiones:

> Un 73 % de los abogados esperan integrar la IA generativa en su trabajo jurídico en los próximos 12 meses.
>
> No existe un consenso sobre si la IA generativa supone una oportunidad o una amenaza.
>
> Casi tres cuartas partes de los abogados afirman saber cómo aplicar la IA generativa a su trabajo.

La empresa afirma que este tipo de respuestas no se obtuvo en las cinco ocasiones anteriores en las que tuvo lugar la encuesta:

> En 2019, por ejemplo, la IA tradicional seguía siendo una tecnología que los abogados vislumbraban en el horizonte, y más de la mitad (58 %) predecían que acabaría repercutiendo en su trabajo en los tres años siguientes. En 2023, en cambio, observamos que un apabullante 73 % de los abogados esperan integrarla en su trabajo en el próximo año.

Los resultados son de interés porque indican que, como hemos visto, existe un numeroso grupo de abogados, que prestan su trabajo tanto en Europa como en Estados Unidos y el Reino Unido, que es consciente, en mayor o menor medida, de las posibilidades que ofrece a su trabajo el uso de programas de IA.

2.2) Aranzadi hace en su página web (https:// www. aranzadilaley.es/inteligencia-artificial) la siguiente declaración:

> Desde *Aranzadi LA LEY* llevamos más de cinco años trabajando con Inteligencia Artificial y machine learning en nuestros servicios. Apostamos por la IA Generativa garantizando la *seguridad, fiabilidad y confidencialidad* a través de soluciones que te ofrecen respuestas basadas en *contenido y conocimiento rigurosamente actualizado y en entornos 100% seguros***.**

En la misma página, más adelante, se concreta lo siguiente:

> Sabemos que cualquier decisión crítica en el ámbito profesional legal ha de ser constatada con fuentes fiables y adoptarse con la máxima seguridad jurídica. Nuestra *gran base de conocimiento* constantemente actualizada es la piedra angular, y lo que nos permite acelerar el desarrollo de nuevas *soluciones de vanguardia y plataformas y servicios personalizados***.**
>
> Somos un *proveedor LegalTech* que aplica rigurosamente la IA generativa para ayudar a los profesionales jurídicos a trabajar de forma *más eficiente*, contribuyendo a transformar de una manera decisiva la forma de operar de los profesionales jurídicos.

Observamos la voluntad de la empresa de utilizar la IA generativa manteniendo ciertas precauciones: garantizar la "seguridad, fiabilidad y confidencialidad**"** de la documentación jurídica accedida, a través de soluciones que ofrezcan respuestas basadas en **"**contenido y conocimiento rigurosamente actualizado y en entornos 100 % seguros".

Esta información es importante porque implica que la empresa quiere dar garantías sobre el acceso a documentación existente y actualizada por la misma empresa evitando que su sistema dé respuestas no confiables.

2.3) La empresa Lefebrve (2024) ha publicado un libro electrónico sobre *Regulación de la IA y la protección de datos en España* (7).

En este libro (Lefebvre, 2024, p. 12 s), la empresa expresa cuál es el funcionamiento que tiene el programa, denominado GenIA-L, de IA generativa, que ella ha creado. Se indica que el programa genera automáticamente preguntas de ayuda para profundizar en el caso planteado por el usuario o explorar nuevas líneas de investigación.

Una vez establecidas las preguntas, el programa da la respuesta incluyendo las respuestas que la empresa contiene en sus compilaciones o Mementos, que son manuales prácticos de consulta que incluyen en un solo volumen toda la información jurídica de la materia sobre la que versa, y que permiten un acceso rápido y fácil a la información gracias a su tabla alfabética que remite directamente al párrafo que contiene la solución. El contenido de los Mementos asociados a la base de datos de Lefebvre es exactamente el mismo que el de los Mementos en papel.

Este contenido permite la navegación a otros apartados del propio Memento o de Mementos relacionados, así como la navegación a Legislación, Jurisprudencia, Doctrina y cualquier otro documento de la base de datos.

Como puede observarse la empresa Lefebvre explica el modo de usar la IA generativa que integra su programa GenIA-L con la información sobre distintos problemas jurídicos, ordenados atendiendo a categorías jurídicas de la ciencia del Derecho de carácter continental e integrados por los contenidos de las respectivas las fuentes jurídicas de este sistema (ley, jurisprudencia y doctrina) que publica en formato digital y papel.

2.4) El Grupo Editorial Tirant lo Blanch propone como sistema de IA el denominado Sofia, que considera es un asistente jurídico de IA (https://prime.tirant.com/es/sofia-2/).

La página web expresa lo siguiente:

> Sofía supone un cambio de paradigma en la forma de trabajar. Mientras los usuarios escriben o cuando suben un documento ya redactado, SOFÍA lee y comprende. Analiza el ámbito de aplicación y localiza los conceptos legales en los que está trabajando el usuario para ofrecerle, como sugerencias, toda la legislación, formularios, jurisprudencia y doctrina relacionada, pertinentes para su caso.

Esto es así porque

> SOFÍA está diseñada para entender el lenguaje natural y para estar en constante actualización mientras el usuario escribe, borra o selecciona un apartado concreto.
>
> Sofía trabaja con el conocimiento que le proporciona la base de datos de Tirant Prime y con la ayuda de un equipo de expertos juristas que constantemente ajusta sus algoritmos para precisar sus resultados. Por eso, Sofía aprende cada día para que la inteligencia artificial sea más útil para nuestros usuarios.
>
> Sofía comprende el lenguaje natural, no necesita que el usuario se adapte a un formato u orden documental especiales.
>
> Se adapta a las modificaciones sobre los textos (añadidos y eliminaciones) y recalcula sus sugerencias en segundos. Y además, es capaz de analizar solo un apartado concreto seleccionado dentro de un documento más amplio.
>
> Como una solución separada, integrada en la base de datos o integrada en el gestor de despachos, Sofía se adapta al espacio de trabajo del abogado para con un objetivo único: cambiar el paradigma de la búsqueda documental. Gracias al entendimiento de Sofía, los abogados no tendrán que buscar documentos, sino aceptar o rechazar las sugerencias de Sofía.
>
> Sofía está a su lado para ayudarle: mientras redacta, mientras lee un documento, si quiere comprobar un escrito que ya tenía en su ordenador… Siempre, utilizando la I.A. jurídica para sugerirle los documentos más pertinentes y siempre, sin reemplazar ni interferir en el trabajo del abogado.

Por tanto cabe concluir que en el caso de Tirant lo Blanch existe un sistema de IA (SOFIA) que comprende el lenguaje conceptual jurídico español cuando lo introduce el usuario al preguntar al sistema por un caso y le da acceso a la información jurídica, de carácter continental, que almacena en forma actualizada la base de datos jurídica de la empresa denominada Tirant Prime.

2.5) La empresa V-Lex ha creado su programa de IA denominado Vincent AI. Sobre sus características y funciones en la página web de la empresa se dice lo siguiente (https://vlex.es/products/vincent-ai):

> Vincent AI es el primer asistente de búsqueda que convierte la tradicional búsqueda por palabras en algo completamente nuevo. Mejora tu eficiencia y tu visión legal como nunca antes lo habías hecho.
>
> AI combinar el comportamiento de búsqueda natural con la velocidad de lectura, comprensión y búsqueda mediante IA generativa y tecnología inteligente, Vincent AI puede ahorrarte tiempo, mejorar la precisión de tu investigación y aumentar tu productividad y eficacia, lo que la convierte en una herramienta imprescindible para cualquier profesional del Derecho.
>
> Vincent AI responde a preguntas jurídicas debidamente fundamentadas en la legislación, jurisprudencia y doctrina. Vincent AI te ayuda a comprender el contexto de la respuesta con resúmenes de las fuentes utilizadas y su acceso directo.
>
> A partir de un determinado argumento, Vincent AI puede argumentar a favor o en contra, usando la legislación, la jurisprudencia y la doctrina como base.

> El usuario puede descartar aquellos contenidos que no crea relevantes para su caso y que Vincent AI los haya incorporado en su análisis inicial. Este nivel de control no tiene precedentes.
>
> Vincent AI es capaz de proporcionar una visión paralela sobre la misma cuestión jurídica en múltiples jurisdicciones para una mejor investigación de derecho comparado. Esta nueva y potente función te permitirá distinguir claramente los puntos de derecho de las áreas de interés para tu empresa o cliente.
>
> Vincent AI leerá y extraerá los conceptos jurídicos clave de una sentencia para proporcionarte una comprensión de las cuestiones que se abordan en el caso.
>
> Además, dado que en muchas regiones del mundo solo se notifican alrededor del 20 % de los casos, Vincent AI puede ayudarle a encontrar el caso adecuado mediante resúmenes generados automáticamente que te ayudarán a comprender el caso sin necesidad de leer la sentencia completa.
>
> La intuitiva interfaz "arrastrar y soltar" de Vincent permite a los investigadores cargar cualquier documento pertinente, ya sea un caso, un esqueleto argumental, un contrato, una ley o cualquier otro tipo de documento jurídico. También se puede teclear o pegar la información en el cuadro de texto. Vincent devolverá entonces un conjunto único de resultados de búsqueda de gran relevancia, un proceso que habría llevado horas compilar manualmente.
>
> Vincent analiza la jurisprudencia y legislación dentro del texto y recomienda contenidos relacionados con tu investigación jurídica que están disponibles en vLex, incluso si no se han mencionado en el documento. […]
>
> Vincent AI ha sido diseñado para facilitar la máxima transparencia al usuario, rompiendo el efecto “caja negra” que en muchos casos se da en el uso de herramientas similares. Para ello vLex ha diseñado un sistema basado en la trazabilidad del contenido, de tal forma que el usuario puede ver qué contenidos o parte de los mismos son usados como base para la argumentación.
>
> Para prevenir “alucinaciones” y generación de respuestas falsas, asociadas frecuentemente al uso de modelos LLM, vLex ha diseñado un sistema que únicamente procesa información y conocimiento que se encuentre en su base de datos, de manera que se garantiza la certeza y la actualización permanente del contenido con el que se generan las respuestas.
>
> Busca entre millones de contenidos. Contenido global, exclusivo y siempre actualizado. La mejor garantía es el uso exclusivo de contenido propio de vLex, exhaustivo, autoritativo y siempre actualizado.

Ilustra sobre el detallado funcionamiento del sistema que hace la página web una presentación sobre el mismo que está publicada en Youtube (8).

Como ha podido observarse VLex explica el funcionamiento de su sistema de acceso a documentación jurídica de carácter continental, utilizando su herramienta de IA generativa denominada Vincent AI. Pone énfasis en que sus sistemas sólo utilizan la información que está en su sistema de base de datos jurídica a efectos de dar seguridad al sistema y evitar “alucinaciones”, es decir respuestas falsas, y sin sentido ni relación con lo que el usuario requiere al sistema, que se producen en el uso de IA.

Puede verse con respecto a la expresión “alucinación” en relación a IA la definición de Ji Ziwei (Ziwei y otros, 2023, p. 31):

> Hallucination is an artifact of Natural Language Generation and is of concern because they appear fluent and can therefore mislead users. In some scenarios and tasks, hallucination can cause harm.

El tema de las alucinaciones es relevante en el ámbito jurídico. Ello es así porque las respuestas falsas en relación a una materia como el Derecho son posibles. Mucho más si no hay conciencia en el usuario de las diferencias culturales en relación al Derecho Common law o al Derecho continental. También lo son si el usuario no tiene conciencia con respecto a las características del lenguaje jurídico del idioma que usa.

Todo ello es más relevante en el caso de V-Lex porque su sistema ofrece la posibilidad de conocer el derecho comparado en relación a la materia sobre la que versa la pregunta del jurista que inicialmente sea hecha, en nuestro caso, en relación al Derecho español y luego se quiera comparar la argumentación jurídica propuesta con la que se ofrece para casos similares por el Derecho propio de varios países. Ha de tenerse en cuenta que la empresa V-Lex, en colaboración con otras empresas de bases de datos jurídicas, oferta en su sistema el acceso a bases de datos de más de cien países de todos los continentes, que tienen lenguajes y sistemas jurídicos diferentes: más allá del Derecho continental o el Common law.

4.3.2.2. Procesos judiciales concretos

En lo referido a la realización de procesos las iniciativas adoptadas son las que se expresan a continuación.

1) Aranzadi avanza algo más que lo hasta aquí expresado en su aplicación de la IA. Se refiere en concreto a que la IA ya se puede usar en el auxilio a la realización de procesos judiciales.

Esto se observa al considerar las funciones encomendadas a dos aplicaciones que la empresa ofrece a los interesados bajo las denominacio-

nes: Aranzadi Fusión (https://www. aranzadilaley.es/productos/aranzadi-fusion.html) y Aranzadi One (https://www.aranzadilaley.es/productos/aranzadi-one.html).

Así se puede ver en las notas más relevantes de ambos productos recogidas en sus respectivas páginas web.

> Aranzadi Fusión es el sistema de gestión de despachos para el control y seguimiento de todos los asuntos, que te permitirá: estandarizar procesos y mejorar la rentabilidad de tu negocio, maximizando la seguridad y el servicio que prestas a tus clientes. Todo ello, potenciado con la información jurídica más completa y especializada del mercado,

También se indica:

> […] es el primer ecosistema legal en incorporar una capa de Inteligencia Artificial con machine learning en los procesos de gestión de Notificaciones Judiciales. La IA que aplicamos es capaz de leer y extractar la información relevante de las Notificaciones Judiciales y sugerir acciones de gestión, como es la asociación con el expediente correcto, la incorporación de una anotación en la agenda, etc. Esta ventaja supone un ahorro considerable de tiempo en la revisión y gestión de las notificaciones de los asuntos judiciales.

Con respecto a "Aranzadi One", la página web expresa:

> Aranzadi One. Base de datos jurídica + software de gestión, […] La manera más sencilla de agilizar el día a día de los despachos pequeños y abogados autónomos. […] Una solución integral que incorpora un software de gestión de asuntos para la realización de todas las tareas administrativas de tu despacho, y una base de datos jurídica de información legal fiable.

La página web continúa en otro apartado:

> Aranzadi One, ahora más ágil y sencillo, incorporando una capa de Inteligencia Artificial para la automatización de la gestión de las notificaciones judiciales.
>
> Aranzadi One da un nuevo paso más allá, e incorpora un nuevo módulo con capacidades de Inteligencia Artificial para que ahorres tiempo y dinero en la gestión de las notificaciones judiciales.

Con estos productos concretos la empresa Aranzadi se compromete a utilizar la IA, construida especialmente usando el procedimiento de aprendizaje de máquina, al: 1) proporcionar métodos para organizar la actividad profesional propia del despacho de abogados, y 2) organizar la gestión de las notificaciones judiciales que se emiten en forma automatizada.

2) La Ley / Wolters Kluwer, también oferta una solución informática desarrollada bajo el nombre de "Legisway", dirigida a abogados, en la que se utilizan herramientas de IA que auxilian a la realización de los procesos judiciales. Sobre las funciones del programa "Legisway". Wolters Kluwer en su página web (https:// www.wolterskluwer.com/es-es/solutions/legiway) indica que:

> Legisway… Asume el control de tu información jurídica y acelera el rendimiento de tu empresa con el software de gestión jurídica todo en uno de Wolters Kluwer […] Legisway combina experiencia jurídica y de software para aportar información del mundo real a soluciones tecnológicas punteras. Nuestras soluciones todo en uno permiten a los profesionales jurídicos aumentar la eficiencia y la colaboración con el resto de la empresa en aras del crecimiento empresarial.

Un módulo del programa (el denominado Legisway Analyzer) está construido usando IA:

> la IA integrada de Legisway, ayuda a los departamentos legales a acelerar la revisión de contratos al encontrar respuestas a sus preguntas contractuales. Al consultar contratos en todos los idiomas, realizará un seguimiento de los riesgos y obligaciones, y se mantendrá al tanto de los riesgos legales, de reputación y comerciales que puedan afectar al negocio.

Estos sistemas auxilian a la gestión de los despachos de abogados y a estudiar, en este caso con IA, los contratos de los que los despachos tengan que ocuparse observando los riesgos legales y no legales (económicos, especialmente) que puede acarrear su puesta en práctica por quienes los firmen o acepten.

4.3.2.3 Consideraciones generales

Hemos visto aquí que la estrategia adoptada, generalmente, en España en el uso de IA es la de añadir al acceso que los sistemas proporcionan tradicionalmente a las bases de datos jurídicas, las posibilidades que ofrecen los programas de IA: el aprendizaje de máquina, el procesamiento del lenguaje natural, en ocasiones incluso ChatGPT.

En otros casos se accede, en forma dialogada: mediante argumentaciones, a documentación jurídica que pueda basar la posición jurídica más adecuada a utilizar por el abogado que se ocupa del asunto concreto por el que se realiza la consulta.

Cabe observar aquí que las propuestas indicadas no llegan a ser tan específicas como las que hacen las empresas que se ocupan del Common law, cuya función ya hemos indicado (supra en 4.3.1) está concretada en mayor medida, bien en relación a la propuesta y revisión de contratos o al auxilio al procedimiento judicial.

Esto es así porque, como hemos indicado con anterioridad, las seis propuestas recogidas en el

apartado sobre “Aspectos generales”, hechas por el CENDOJ y las empresas señaladas, parecen estar orientadas a otro objetivo que las que se dirigen al ámbito de Common law.

Las propuestas españolas declaran que toda su oferta de documentación se realizará con auxilio de distintas técnicas de IA, incluyendo, en varios casos, las propuestas de la “IA Generativa”.

Al respecto hay que decir que esta actitud generalizada incrementa notablemente la dificultad de hacer uso de estas técnicas porque no es lo mismo diseñar programas sobre casos y problemas concretos, que construir programas con fines generales para solventar, con todo el Derecho vigente en un país de Derecho continental, todo tipo de problemas. Como hemos visto este es el objetivo manifestado por la mayor parte de testimonios recogidos.

Está, en cambio, más limitado el objetivo, en definitiva: se cuenta con mayor concreción, en las propuestas situadas en el apartado titulado “Realización de procesos”. Estas propuestas ofertan programas ya desarrollados y que pueden ser adquiridos en el mercado bajo la denominación Aranzadi Fusión, Aranzadi One y Legisway Analyzer (Wolters Kluwer).

Ha de sumar a lo anterior que el último programa (Legisway Analyzer) tiene un cometido con relación a contratos que conecta con el que recogíamos con respecto a experiencias que suceden en países de Common law.

A lo anterior hay que añadir que, en todo caso, para valorar el estado de las meritorias iniciativas españolas indicadas es preciso tener presente que, tanto en la temática acceso a documentación jurídica, como en otras que pueda considerarse, es inevitable tener conciencia de que los resultados alcanzables por las aplicaciones o programas señalados siempre han de estar limitados, porque en general lo han de estar los sistemas, programas o aplicaciones de IA que se hagan en español. Ello es así porque el desarrollo del procesamiento de lenguaje natural del idioma español no está estudiado e implantado en un grado similar al hecho en relación al idioma inglés. Y ello es porque, como dice Asunción Gómez-Pérez, experta en IA y procesamiento del español y académica de la Real Academia Española (Gómez-Perez, 2023, p. 97):

> […] en esta carrera tecnológica que persigue obtener modelos más ricos y productivos, el español progresa detrás del inglés. No existe todavía una metodología sistemática para evaluar modelos de lenguaje en español para diferentes tareas. Se convierte, pues, en [un] desafío el desarrollo de una metodología de evaluación comparativa de los modelos de lenguaje de propósito general y específico en nuestra lengua.

También hay que seguir insistiendo en que el sistema jurídico español es en español y en formato continental, es decir está regido por el principio de que el Derecho de referencia primordial para cualquier actividad es el contenido en las leyes, mientras que el de los países de Common law es el jurisprudencial: las sentencias judiciales. Ello hace que los dos sistemas jurídicos tengan que atender a regímenes de fuentes jurídicas diferentes: de mayor heterogeneidad en los países de Derecho civil o continental que en los de Common law.

Todo lo anterior —que en parte es positivo porque indica que, al menos, el lenguaje jurídico está más acotado que el lenguaje general: es un uso de lenguaje específico— no puede hacernos olvidar que ello no deja de incrementar las dificultades para construir sistemas de IA de acceso a documentación jurídica en España, tal y como sucede en otros lugares porque, obviamente, estas dificultades son comunes a las existentes en países de habla no inglesa cuya cultura jurídica es continental. Es un hecho que en estos países no sirven las propuestas hechas desde otros sistemas jurídicos, tal y como expresa el profesor finlandés Ahti Saarenpää (2024, p. 39):

> Legal concepts are always related to some legal informational environment. We are talking about systems and systems thinking. Crossing system boundaries easily leads to an incorrect legal view. Legal principles in force and their limits are not recognised.

En este momento ya conviene completar lo dicho exponiendo el contenido de la regulación europea sobre IA en lo referido a aplicaciones auxiliares de actividades jurídicas que utilicen dicha tecnología.

5. El Reglamento europeo de inteligencia artificial y el uso de la IA en el acceso a textos jurídicos y en la realización de actividades judiciales

Aquí exponemos, sucintamente, el contenido del Reglamento europeo sobre inteligencia artificial, en lo referido al auxilio de las actividades jurídicas por la IA. Ha de tenerse en cuenta que este Reglamento ha sido aprobado por el Consejo de la Unión Europea el 21 de mayo de 2024, tras haberlo sido por el Parlamento Europeo el 13 de marzo de 2024. Aquí seguimos el articulado del texto aprobado en esta fecha (9).

La norma establece que, una vez promulgado, el Reglamento de Inteligencia Artificial debe aplicarse dos años después de su entrada en vigor.

Lo primero a consignar aquí es que la norma tiene alcance general: está dirigida a regular el uso de la inteligencia artificial en todo tipo de actividades. Es decir, introduce un marco normativo común para la inteligencia artificial. O lo que es lo mismo: su contenido y su ámbito de aplicación abarca a todos los sectores y tipos de programas de inteligencia artificial.

Estudiando su contenido cabe decir que sus objetivos son los dos siguientes (considerando 1 del Reglamento):

1. Garantizar que los sistemas de Inteligencia Artificial utilizados en la Unión Europea e introducidos en el mercado europeo, sean seguros, respeten los derechos de los ciudadanos, y eviten las afectaciones de los sistemas a las que hemos hecho mención al considerar casos de uso de la IA en actividades jurídicas.
2. Estimular la inversión y la innovación en el ámbito de la IA en Europa.

Aquí nos vamos a fijar únicamente en las disposiciones que en la norma se hacen con respecto a la materia que ha sido estudiada en este trabajo: el desarrollo de aplicaciones de inteligencia artificial en las actividades jurídicas. Se hacen consideraciones al respecto en los siguientes apartados: riesgo en las aplicaciones jurídicas: los sistemas de IA de alto riesgo (apartado 5.1); la promoción de los sistemas de acceso a documentación jurídica que usan de IA (apartado 5.2), y reglas e indicaciones a considerar en el desarrollo y uso de los sistemas de IA generativa según el Reglamento (apartado 5.3).

5.1. Riesgo en las aplicaciones jurídicas: sistemas de IA de alto riesgo

El Reglamento establece (art. 6: 1) que

> […] un sistema de IA se considerará de alto riesgo cuando reúna las dos condiciones que se indican a continuación:
>
> a) que el sistema de IA esté destinado a ser utilizado como componente de seguridad de un producto que entre en el ámbito de aplicación de los actos legislativos de armonización de la Unión enumerados en el anexo I, o *que el propio sistema de IA* sea uno de dichos productos; y
>
> b) que el producto del que el sistema de IA sea componente de seguridad *con arreglo a la letra a)*, o el propio sistema de IA como producto, deba someterse a una evaluación de la conformidad realizada por un organismo independiente para su introducción en el mercado o puesta en servicio con arreglo a los actos legislativos de armonización de la Unión enumerados en el anexo I.

Además de lo dicho el mismo artículo 6 en su número 2, da cuenta de que

> […] se considerarán de alto riesgo los sistemas de IA a que se refiere el Anexo III.

A nuestros efectos lo que importa consignar aquí es que el Anexo III, 6 y 8, incluye a los siguientes sistemas en la categoría de sistemas de alto riesgo:

- Anexo III.6. Aplicación de la ley: Sistemas de IA utilizados para evaluar el riesgo de una persona de convertirse en víctima de un delito. Polígrafos. Evaluación de la fiabilidad de las pruebas durante investigaciones o procesos penales. Evaluación del riesgo de delincuencia o reincidencia de una persona no basada únicamente en la elaboración de perfiles o en la evaluación de rasgos de personalidad o comportamientos delictivos anteriores. Elaboración de perfiles durante detecciones, investigaciones o enjuiciamientos penales, y
- Anexo III.8. Administración de justicia y procesos democráticos: Sistemas de IA utilizados en la investigación e interpretación de hechos y en la aplicación de la ley a hechos concretos o utilizados en la resolución alternativa de conflictos. Influencia en los resultados de elecciones y referendos o en el comportamiento electoral, excluidos los productos que no interactúan directamente con las personas, como las herramientas utilizadas para organizar, optimizar y estructurar campañas políticas.

Esto significa que el Reglamento prescribe que el auxilio a las actuaciones judiciales por medio de sistemas de IA está considerado, en congruencia con la trascendencia que tienen dichas actividades, como propio de sistemas de alto riesgo. Lo que requiere que en su tramitación y puesta en funcionamiento se atienda a los requerimientos especiales que son demandados a este tipo de sistemas de alto riesgo que la norma estipula, o lo que es lo mismo la intervención de los diferentes organismos europeos y nacionales que la norma prevé se implanten en los respectivos ámbitos de actuación.

5.2. Promoción de los sistemas de acceso a documentación jurídica

El artículo 6 del Reglamento establece actividades para las que no se precisa la puesta en acción de los mecanismos previstos para los sistemas de alto riesgo.

El artículo 6, a la vez que regula normas de clasificación de los sistemas de IA de alto riesgo, excepciona a otros de la consideración de sistemas de dicha calificación. Se trata del apartado 2a del artículo:

> 2a. No obstante lo dispuesto en el apartado 2, los sistemas de IA no se considerarán de alto riesgo si

> no suponen un riesgo significativo de daño para la salud, la seguridad o los derechos fundamentales de las personas físicas, incluido el hecho de no influir materialmente en el resultado de la toma de decisiones. Este será el caso si se cumplen uno o varios de los siguientes criterios:
>
> (a) el sistema de IA está destinado a realizar una tarea de procedimiento limitada;
>
> (b) el sistema de IA está destinado a mejorar el resultado de una actividad humana previamente realizada;
>
> (c) el sistema de IA está destinado a detectar patrones de toma de decisiones o desviaciones de patrones de toma de decisiones anteriores y no está destinado a sustituir o influir en la evaluación humana previamente completada, sin la debida revisión humana; o bien
>
> (d) el sistema de IA está destinado a realizar una tarea preparatoria de una evaluación pertinente a efectos de los casos de uso enumerados en el anexo III. No obstante lo dispuesto en el párrafo primero del presente apartado, un sistema de IA se considerará siempre de alto riesgo si el sistema de IA realiza la elaboración de perfiles de personas físicas.

Tal y como hemos explicado con anterioridad al mostrar las virtualidades de la IA para el acceso a textos jurídicos (apartado 4), estas excepciones recaen directamente en las funciones propias de dichos sistemas. Ha de destacarse que son varios de los subapartados que contiene el artículo 6, 2a los que son de aplicación para dichos sistemas.

5.3. El desarrollo y uso de los sistemas de IA generativa en el Reglamento

El uso de la IA generativa al que hemos hecho referencia al exponer varios ejemplos de las funciones a satisfacer por los sistemas de IA en cuanto herramientas de acceso a documentación jurídica (apartado 4), requiere que hagamos alguna consideración sobre la regulación de la IA generativa en el Reglamento.

Lo primero que hay que decir es que si bien el Reglamento europeo no menciona explícitamente a los sistemas de IA generativa, si hay un considerando del Preámbulo de la norma que los menciona. También existen varios artículos que contienen algunas implicaciones para los sistemas de IA generativa.

En relación al Preámbulo, en el considerando número 99 se dice:

> (99) Los grandes modelos de IA generativa son un ejemplo típico de un modelo de IA de uso general, ya que permiten la generación flexible de contenidos, por ejemplo, en formato de texto, audio, imágenes o vídeo, que pueden adaptarse fácilmente a una amplia gama de tareas diferenciadas.

En coherencia con el contenido de este considerando cabe mencionar algunos artículos del Reglamento que deben tener en cuenta los fabricantes, diseñadores y usuarios de estos sistemas. Son, por ejemplo, los siguientes.

- Artículo 1 - Objeto y ámbito de aplicación: El artículo establece que la ley se aplica a todos los sistemas de IA, independientemente de la tecnología utilizada. Esto significa que la IA generativa está dentro del alcance de la ley.
- Artículo 3 - Definiciones:

  > "sistema de IA": un sistema basado en una máquina diseñado para funcionar con distintos niveles de autonomía, que puede mostrar capacidad de adaptación tras el despliegue y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar información de salida, como predicciones, contenidos, recomendaciones o decisiones, que puede influir en entornos físicos o virtuales.

  La IA generativa satisface esta definición.
- Artículo 4 - Requisitos generales:

  > Los proveedores y responsables del despliegue de sistemas de IA adoptarán medidas para garantizar que, en la mayor medida posible, su personal y demás personas que se encarguen en su nombre del funcionamiento y la utilización de sistemas de IA tengan un nivel suficiente de alfabetización en materia de IA, teniendo en cuenta sus conocimientos técnicos, su experiencia, su educación y su formación, así como el contexto previsto de uso de los sistemas de IA y las personas o los grupos de personas en que se utilizarán dichos sistemas.

  Esto implica que proveedores y responsables de la introducción y uso de los sistemas de IA generativa, como los de cualquier otro sistema, han de instruir adecuadamente sobre las características y consecuencias de uso de este tipo de aplicaciones de las tecnologías de la información y la comunicación.
- Consideraciones adicionales: El Reglamento también establece, en varios artículos y apartados que los datos utilizados para entrenar sistemas de IA deben ser recogidos y utilizados de manera responsable, de acuerdo con las normas de protección de datos de la Unión Europea. Obviamente ello también es de aplicación en el caso de la IA generativa.

Lo mismo sucede en la regulación que se prevé con respecto a que la Comisión Europea tiene la

potestad de adoptar actos delegados para especificar los requisitos técnicos para los sistemas de IA de alto riesgo. Es posible que estos actos delegados tengan un impacto en IA generativa en el futuro.

En resumen, si bien la Ley Europea de IA no menciona directamente a la IA generativa, existen varios artículos que tienen implicaciones para este modelo de lenguaje. Por ello los desarrolladores de IA generativa deben ser conscientes de los requisitos y adoptar las medidas necesarias para cumplir con el Reglamento.

Completando lo anterior, se presenta en el siguiente apartado varias opiniones de juristas que se manifiestan en relación al uso de la inteligencia artificial por una actividad jurídica diferente a la del acceso a textos jurídicos: la realización de la aplicación del Derecho por los jueces.

6. El uso de la inteligencia artificial por las actividades judiciales

Distinguimos en posiciones a favor y críticas del uso de herramientas de IA en las actividades propias de los profesionales del Derecho, especialmente los jueces. Esto lo hacemos aquí exponiendo, resumidamente, algunas posiciones que se manifiestan tanto en la cultura de Common law como en la del Derecho continental, a favor de dicho auxilio (6.1), y, en otras ocasiones, en sentido crítico hacia él (6.2). Al final se aportará unas consideraciones generales sobre lo expuesto en el apartado (6.3).

6.1. Posiciones a favor

En coherencia con el fuerte interés de los abogados por la IA que mostraba la encuesta internacional de la empresa Wolters Kluwer hecha en 2023 y que mencionábamos más arriba, son numerosos los testimonios que cabe encontrar en la literatura, e incluso en las normas, sobre la relevancia e interés del uso de la IA en otras actividades jurídicas, distintas a la del acceso a textos jurídicos.

A continuación, aportamos tres testimonios sobre la materia. Los dos primeros se refieren a propuestas referidas al auxilio, e incluso reemplazo, de la actividad judicial mediante la apelación a la implantación del "juez robot". El tercer testimonio, de carácter normativo, modera las previsiones hechas en los dos primeros sin dejar de admitir la relevancia que tiene la IA para incrementar la eficiencia de los procedimientos y actividades jurídicas.

1) Una propuesta en el sentido expresado: proponer la IA para cualquier actividad jurídica, incluso para reemplazar la propia del juez en el caso de la aplicación judicial del Derecho, es la que hace Georgios Zekos, abogado y economista griego, en su trabajo titulado, significativamente, *Advanced Artificial Intelligence and Robo-Justice* (Zekos, 2022), fijándose en experiencias y regulaciones propias de países de Common law y Derecho continental.

Zekos expone en la obra citada que la tecnología digital está transformando el panorama de la resolución de conflictos jurídicos mediante la expansión y transición del fenómeno denominado métodos "alternativos" de resolución de conflictos (en inglés "ADR"), al de métodos de resolución "online" de conflictos (en inglés "ODR") y, gracias a las propuestas de desarrollo de la IA avanzada, a la progresiva implantación del juez robot *(Ibidem,* p. 412). Es decir, del juez máquina como sustituto del juez persona.

A estos efectos Zekos en su obra analiza la aplicación de la IA en el ámbito jurídico y muestra las seguras, en su opinión, perspectivas futuras de la implantación de una justicia robótica en una sociedad situada en la era de la inteligencia artificial avanzada. Para ello explora el sistema de justicia actual y la influencia de la IA en la práctica de los tribunales y el arbitraje. Examina, también, el papel transformador de la IA en todos los ámbitos jurídicos en temas como la personalidad jurídica y la responsabilidad de la IA. Este análisis muestra, según el autor *(Ibidem,* p. 417-420), que la tecnología digital está generando un número cada vez mayor de disputas y, al mismo tiempo, está desafiando la efectividad y el alcance de las vías tradicionales de resolución de disputas, mostrando la necesidad de la generación de un nuevo sistema de justicia robótica.

2) En España ha hecho una aportación referida a la posible implantación del juez robot Carolina Sanchis, profesora de Derecho Procesal, quien, en su trabajo titulado *Inteligencia artificial y decisiones judiciales: crónica de una transformación anunciada,* recogiendo opiniones coincidentes de otros autores españoles sobre la materia, expresa lo siguiente (Sanchis-Crespo, 2023, p. 82):

> Tras haber examinado […] la capacidad del juez robot para cumplir la función jurisdiccional en su vertiente de declarar el derecho debemos concluir que puede hacerlo, pero con las limitaciones que se han indicado.
>
> Estas restricciones podrán irse difuminando conforme avance la tecnología quedando siempre las cuestiones más complejas sometidas al conocimiento de los jueces humanos.
>
> En cuanto a la conveniencia de que el juez robot ejercite esa función jurisdiccional, desde el punto de vista de la imparcialidad y la independencia que son sólo dos de los obstáculos a considerar, la conclusión es también positiva […].

3) Desde esa posición en la que se considera que el juez robot puede cumplir al menos una parte de la función jurisdiccional y en la que se juzga que es conveniente que lo haga, cabe plantear otras reflexiones mirando al futuro que establece el Decreto Ley español sobre eficiencia en la Justicia y su referencia a la inteligencia artificial.

El Decreto Ley ya está aprobado y refrendado por las Cortes. Es el mencionado más arriba (apartado 3, al final). A continuación nos fijamos en el contenido de la regulación en la que se menciona la IA en la que, como vamos a ver, inicialmente, no se hace mención al juez robot.

Hablamos del *Real Decreto-ley 6/2023, de 19 de diciembre, por el que se aprueban medidas urgentes para la ejecución del Plan de Recuperación, Transformación y Resiliencia en materia de servicio público de justicia, función pública, régimen local y mecenazgo* (Real Decreto-ley 2023). En la norma observamos que se hace referencia a la IA en el ámbito judicial, pero cuando se menciona se hace en estos términos:

> TÍTULO III
>
> De la tramitación electrónica de los procedimientos judiciales
>
> CAPÍTULO II
>
> Tramitación orientada al dato
>
> Artículo 35. Principio general de orientación al dato.
>
> 1. Todos los sistemas de información y comunicación que se utilicen en el ámbito de la Administración de Justicia, incluso para finalidades de apoyo a las de carácter gubernativo, asegurarán la entrada, incorporación y tratamiento de la información en forma de metadatos, conforme a esquemas comunes, y en modelos de datos comunes e interoperables que posibiliten, simplifiquen y favorezcan los siguientes fines:
>
> […]
>
> k) La aplicación de técnicas de inteligencia artificial para los fines anteriores u otros que sirvan de apoyo a la función jurisdiccional, a la tramitación, en su caso, de procedimientos judiciales, y a la definición y ejecución de políticas públicas relativas a la Administración de Justicia.

Como puede observarse el texto jurídico no menciona, como en los ejemplos anteriores, al "juez robot" sino que se limita a prescribir reglas técnicas a satisfacer por los programas de información y comunicación de IA *"que sirvan de apoyo a la función jurisdiccional, a la tramitación, en su caso, de procedimientos judiciales".*

Por lo tanto aquí, si bien hemos visto dos ejemplos aportados por juristas que reflexionan sobre la posibilidad del juez robot mediante el auxilio de la IA, en cambio también hemos observado que en la norma que en España regula el auxilio de la IA no se habla explícitamente de dicha posibilidad sino de su posible apoyo a la tramitación de procedimientos judiciales y a la de la función jurisdiccional.

Volvemos sobre el contenido de esta regulación y las críticas que ha suscitado por el Consejo General del Poder Judicial español en el próximo apartado.

6.2. Posiciones críticas

Además de las posiciones que acabamos de expresar, también existen otras posturas que o bien expresan su crítica o su prevención con respecto al uso de las herramientas de IA por los jueces, y otras que critican, incluso, la implantación de ciertas herramientas de IA en la Administración de Justicia y su puesta a disposición del juez.

Para presentar estar posiciones aquí nos referimos a lo siguiente. En primer lugar, hacemos un breve resumen sobre la crítica que se realizó por el Consejo General del Poder Judicial a la regulación sobre IA y actividad judicial que, en su opinión, contiene en verdad el *Real Decreto-ley 6/2023, de 19 de diciembre, por el que se aprueban medidas urgentes […] en materia de servicio público de justicia* (Real Decreto-ley 2023). La crítica se centra, como vamos a ver, en presentar la opinión de que esta regulación da pie a hablar del juez robot. En segundo lugar, aportamos una posición de interés sobre la discusión, elaborada la experiencia del uso de la IA en actuaciones judiciales realizadas en Estados Unidos.

1) El contenido del Real Decreto-ley mencionado en lo referido al servicio público de justicia está recogido en el Libro primero y en sus nueve títulos (uno de ellos de carácter preliminar). El objeto de la norma se limita a, como se denomina el Libro, la organización del servicio público de justicia, siendo escasa la mención explícita a la IA como hemos expresado con anterioridad (al final del apartado 6.1). Así es: el libro primero del texto legal se llama Medidas de Eficiencia Digital y Procesal del Servicio Público de Justicia. De su regulación se ocupan los artículos, que van del 1 al 104.

Lo que aquí nos interesa destacar es el contenido de su título III que versa sobre la tramitación electrónica de los procedimientos judiciales, recogiendo, como antes hemos expresado, en su capítulo II (que contiene los arts. 35 a 38), denominado tramitación orientada al dato, el art. 35, que es el único de la norma que hace mención a la IA.

Tal y como se ha expresado con anterioridad, recogiendo el texto del artículo 35 (párrafo 1 y subpárrafo k), la apelación a la IA es debida a la implantación por la norma del principio general de

orientación al dato, lo que se hace porque se entiende que este principio contribuirá a un mejor diseño de las políticas públicas, gracias al análisis de los numerosos datos que genera la Administración de Justicia, debido a que la tramitación de expedientes dejará de estar orientada al documento y pasará a estar orientada al dato. De ahí antes mencionáramos que la apelación a la IA, considerando en abstracto el art. 35, parece no implica deseo alguno de implantar el juez robot.

Ello no obstante esta última afirmación es cuestionable, como así lo ha expresado el Consejo General del Poder Judicial, si se tiene en cuenta que, además de lo dicho, el capítulo VII (artículos 56, 57 y 58), del mismo título III, está denominado con las expresiones: "De las actuaciones automatizadas, proactivas y asistidas". Además de lo dicho, ha de considerarse que el capítulo VII finaliza el título III que versa sobre la tramitación electrónica de los procedimientos judiciales.

Es por lo anterior que ante el estudio del contenido del capítulo VII y los artículos mencionados del Decreto-ley (arts. 35, 56, 57 y 58), el Consejo General del Poder Judicial en su Pleno del día 24 de febrero de 2022, aprobó su Informe sobre el Anteproyecto de Ley de Eficiencia Digital, que contenía los mismos textos que los del actual Decreto-ley que hemos mencionado.

En su Informe el Consejo del Poder Judicial (CGPJ) se manifestó críticamente en lo referido a las llamadas "actuaciones automatizadas, proactivas y asistidas", producidas por un sistema de información debidamente programado, o en base a datos producidos por algoritmos, en el ámbito de los procedimientos judiciales. La crítica se debía a que el CGPJ advirtió en dicho Informe que si bien la norma establece que las actuaciones automatizadas pueden generarse por defecto a partir de un determinado sistema de información, falta por decir que ello será siempre sin perjuicio de la dirección del proceso, que corresponde a Jueces y Magistrados, que podrán establecer las instrucciones pertinentes sobre su uso o deshabilitación.

A este respecto es de destacar lo que expone el CGPJ en el considerando sexagesimoctavo del informe (Informe, 2022, p. 174):

> La generación por los sistemas de información, con base en algoritmos, de borradores de resolución que contengan determinación de hecho e interpretación del derecho aplicable puede verse como una ayuda o apoyo al ejercicio de la función constitucional de juzgar, pero constituye también, y debe subrayarse, un riesgo para la vigencia del principio de exclusividad jurisdiccional que exige que la tutela de derechos e intereses de los ciudadanos sea prestada exclusivamente por Jueces y Magistrados. En el contexto del uso de técnicas de inteligencia artificial, debe afirmarse que el artículo 24 CE en conexión con el principio de exclusividad jurisdiccional (art. 117.3 CE) garantiza a los ciudadanos el derecho a una resolución fundada en Derecho dictada por un Juez o Tribunal, esto es, el derecho a que su caso sea resuelto por un Juez-persona.

En definitiva, en esta conclusión el CGPJ estima que el Anteproyecto de Ley, hoy ya Decreto-ley, prevé, aunque no lo mencione explícitamente, dar valor a lo generado automáticamente sin contar con la actuación propia de los jueces. El CGPJ reconocería con ello, criticándolo, que en el texto del Decreto-ley se está ante el reconocimiento ilegal, por el Derecho español, del juez robot.

La implantación del juez robot en España está cuestionada también por la doctrina, desde una perspectiva jurídica y técnica, en el trabajo de Elisa Simó y Paolo Rosso (Simó, 2022, pp. 6s) al referirse a "La sustitución algorítmica en la Administración de Justicia" diciendo:

> En el ámbito de la Administración de Justicia este relevo se ve impedido por el principio de exclusividad previsto en los artículos 117.3 de la Constitución Española y 2.1 de la Ley Orgánica del Poder Judicial según el cual el ejercicio de la potestad jurisdiccional en todo tipo de procesos, juzgando y haciendo ejecutar lo juzgado, corresponde exclusivamente a los Juzgados y Tribunales determinados por las leyes… en caso de producirse el reemplazo de los operadores jurídicos por sistemas de IA se trataría de una sustitución supervisada. No va a llevarse a cabo ninguna transformación sin el control de los sistemas de IA por parte de equipos interdisciplinares de expertos. Es decir, la progresiva automatización de los tribunales será implantada de acuerdo con las oportunidades y límites que como sociedad concedamos a la IA.

2) También cabe destacar como opinión crítica a la utilización de la IA en la actuación de los jueces, la expuesta en Estados Unidos por Catherine Forrest, ex jueza federal en Nueva York y abogada en la actualidad. La opinion la ha emitdo en su libro titulado significativamente: *When Machines Can Be Judge, Jury, and Executioner: Justice in the Age of Artificial Intelligence* (Forrest, 2021).

En este libro la autora reflexiona, a partir de su experiencia como jueza, sobre la situación de la administración de justicia en la era de la IA en Estados Unidos. Forrest sostiene que las herramientas actuales de IA utilizadas como instrumento auxiliar en las decisiones judiciales, basan su funcionamiento en concepciones utilitarias de justicia y son inconsistentes con los principios de libertad y justicia individual reflejados en la Constitución y la Declaración de Independencia de los Estados Unidos.

Las afirmaciones las realiza fijándose en el sistema de justicia penal y en las características de los principios que guían los instrumentos de auxilio a todos sus integrantes, a la toma de decisiones, y que ofrece la IA: herramientas que aportan prejuicios sociales como los de raza o sexo. Dice en concreto (Forrest, 2021, p. 1):

> In today's criminal justice system -whether you're aware of it or not- artificial intelligence (AI) is everywhere: it guides carceral sentences, bail decisions, likehoods of arrest, and even the application of autonomous military weapons. Put simply, AI is essential to how we, as an American society, dispense justice.

Tras lo dicho la jueza manifiesta que la herramienta de predicción y evaluación de riesgos y necesidades de los delincuentes ("risk and need assesment for offenders", RNA), de uso habitual en la Administración de Justicia, y que comporta el uso de la IA, influye utilitariamente en las decisiones judiciales, siendo que, en verdad, dichas decisiones han de estar regidas por el principio de libertad frente al de la utilidad.

La influencia utilitaria se debe a que, según su experiencia, el diseño algorítmico de las herramientas de evaluación de riesgos de la IA puede incorporar, por ejemplo, sesgos humanos, y de hecho lo hace. Con lo cual los diseñadores y usuarios de estas herramientas de IA no permiten que exista un grado de compromiso con el ejercicio libre de la justicia o la equidad individual. El problema reside en que las herramientas son diseñadas por empresas privadas, que las crean para otras empresas o las autoridades administrativas, sin que estas conozcan cómo han sido diseñadas y por tanto cómo trabajan o cómo y porqué pueden dar respuestas sesgadas.

Como dice la autora la solución es rediseñar estas herramientas de IA (p. XVI). Ellas:

> […] need a bottom-up redesign. Any redesign must prioritize a theory of justice as fairness throughout the tool. An AI tool designed along these lines […] could have a positive, coursealtering impact on criminal justice reform nationwide.

La autora también indica que en la relación de la Administración de Justicia con la IA hay que actuar como progresivamente se solicita en Estados Unidos para con el diseño de la IA que guía a los "drones" en actuaciones militares (estos robots son considerados "lethal autonomous weapons"). Ello es así porque en Estados Unidos está extendida la opinión de que ese diseño no puede ser hecho por razones utilitarias. La autora indica al respecto (Forrest, 2021, p. 131):

> In contrast to risk and needs assessment tools, a significant amount of attention is being paid to the appropriate design standards and deployment framework for AI tools that can be used to kill us: lethal autonomous weapons. Whether we like it or not, these weapons are being developed by private contractors and state- financed researchers and military personnel all over the world. - In contrast to risk assessment tools, the debate surrounding LAWs [lethal autonomous weapons] is steeped in well-established rules of war, engagement and international humanitarian law.

6.3. Consideraciones generales

Ante la situación expuesta en relación a las posiciones que están a favor del autómata o el juez robot (en 6.1); las que indican que esto no es posible porque quien ha de decidir es el juez según los principios constitucionales; y las posiciones que están en contra del auxilio de la IA a la actividad judicial porque las herramientas de IA no están funcionando correctamente en la administración de justicia al haber sido diseñadas al margen de la satisfacción de los principios de justicia (como indicamos en 6.2), ya se puede entender las diferentes prevenciones y apoyos que establece el Reglamento sobre inteligencia artificial con respecto al uso de la IA en las actividades jurídicas aquí consideradas, a las que nos hemos referido en el apartado 5.

7. Conclusión

Ya podemos dar una respuesta a la pregunta que planteábamos al comienzo de este trabajo (apartado 1):

> ¿el uso de las tecnologías de la información y la comunicación en el acceso a textos jurídicos, a juristas interesados, cambia sus exigencias si se hace realidad el desarrollo de aplicaciones o programas de ordenador denominados de inteligencia artificial para dicha actividad?

A diferencia de lo que ocurre con la actividad de aplicación de los textos jurídicos en los procedimientos judiciales, que ya hemos visto queda muy limitada por el hecho de que el auxilio de la IA a dicha actividad está considerada de alto riesgo y por tanto necesita de las regulaciones normativas europeas y nacionales pertinentes, la actividad de acceso a textos jurídicos realizada con auxilio de la IA resulta claramente potenciada al incrementar y mejorar la calidad de dicha actividad, potenciando la reflexión de los juristas usuarios al aumentar las posibilidades de centrarse en el estudio de los textos jurídicos y poder mejorar las propuestas de solución de los problemas de los que deban ocuparse en su trabajo profesional. Con ello puede mejorar las actividades de todos los juristas una vez que el acceso a textos jurídicos es indispensable en la puesta en acción del Estado de Derecho.

Como en este trabajo ha quedado señalado son varias las iniciativas de las empresas del sector dedicado a proveer el acceso a textos jurídicos que están llevando adelante, desde la perspectiva del Derecho continental, iniciativas destinadas a mejorar las posibilidades que ya ofrecen a las instituciones, organizaciones y empresas, a las que lo facilitan. Es de esperar que la interacción de dichas empresas con actividades de I+D+i que se ocupan en España del desarrollo de la IA en materias básicas como el procesamiento del lenguaje español y el aprendizaje automático multiplicará el efecto de los productos de las empresas del sector.

Notas

(1) Sobre el interés actual entre los juristas españoles por el uso de las Tecnologías de la Información y la Comunicación, e incluso la IA, en su práctica profesional diaria se trata detalladamante, por ejemplo, en la página web: https://www.derechopractico.es/guialegaltech/

(2) Se puede consultar al efecto la página web: https://openai.com/blog/chatgpt-can-now-see-hear -and-speak. Ahí se dice que ChatPGT: "Get instant answers, find creative inspiration, learn something new". O que: "ChatGPT can see, hear, and speak". En realidad estas afirmaciones han de modularse porque: "Cuando se habla de IA no se debe definir como un ente suprahumano, monolítico y abstracto. La realidad de la IA es, hoy en día, mucho menos romántica. Se trata de modelos estadísticos complejos capaces de autoajustarse según reciben nueva información." (Simó, 2022, p. 2).

(3) Queda abierta esta posibilidad en el confuso apartado referido a "Sharing & publication policy" que está recogido en la página web: https://openai.com/policies/sharing-publication-policy. De hecho existen ocasiones en las que las respuestas del sistema está fundamentadas en información contenida en "links" mencionados explícitamente, mientras que en otras ocasiones no se expresa el "link" de información en el que la respuesta está basada.

(4) Se encuentra información sobre el Sistema y sus funciones en la página: https://www. lawgeex.com/. Otros programas de carácter similar comercializados especialmente en países de common law son: 1) el denominado ContractPodAi, que es un sistema de IA que se ocupa de asistir a la creación de contratos, se localiza información al respecto en https://contractpodai.com/; 2) el denominado Legal Sifter, que es un sistema de Inteligencia Artificial para Revisión de Contratos y Cláusulas Contractuales: https:// www.legalsifter.com/; 3) el que lleva por nombre Kira Systems - Inteligencia Artificial para Identificar y Analizar los Contratos: https://kirasystems.com/; 4) el denominado CoCounsel de Casetext - IA para la Revisión de Documentos y Contratos https://casetext.com/; 5) finalmente: eBrevia - Inteligencia Artificial para Análisis de Contratos, ver su función en: https://ebrevia.com/.

(5) Más información se puede encontrar en la página: https://www.luminance.com/. Otros sistemas similares son: 1) Harvey destinado a proporcionar Inteligencia Artificial para Apoyo Legal Integral a Abogados en sus labores de consultoría y litigio, se encuentra información en https://www.harvey.ai/; 2) Lex Machina, que proporciona Inteligencia Artificial para Ayudar a Abogados en el Análisis de Datos de Litigios, hay más información en https://lexmachina.com/.

(6) La encuesta puede obtenerse en: https://www. wolterskluwer.com/es-es/know/futurready-lawyer-2023.

(7) El libro se puede solicitar en: https://lefebvre .es/ tienda/libros-derecho-pdf-gratis/regulacion-de-la-ia-y-la-proteccion -de-datos-en-espana.

(8) https://www.youtube.com/watch?v=hLvqL17eR kE&t=9s.

(9) https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/es/pdf

Referencias

Colombo, Pierre; Pessoa Pires, Telmo; Boudiaf, Malik; Culver, Dominic; Melo, Rui; Corro, Caio; Martins, Andre F. T.; Esposito, Fabrizio; Raposo, Vera Lúcia; Morgado, Sofia; Desa, Michael (2024). SaulLM-7B: A pioneering Large Language Model for Law. arXiv:2403.03883v2 [cs.CL] 7 Mar 2024: https://arxiv.org/abs/2403.03883, 13 páginas.

Consejo General del Poder Judicial (2022). Informe al anteproyecto de ley de eficiencia digital del servicio público de justicia, por la que se transpone al ordenamiento jurídico español la directiva (ue) 2019/1151 del parlamento europeo y del consejo, de 20 de junio de 2019 por la que se modifica la directiva (ue) 2017/1132 en lo que respecta a la utilización de herramientas y procesos digitales en el ámbito del derecho de sociedades.

Consejo General del Poder Judicial (2022). Memoria sobre el estado, funcionamiento y actividades del Consejo general del poder judicial y de los juzgados y tribunales en el año 2021. Secretaría General: Consejo General del Poder Judicial.

CREI. (1983). Gestión automatizada en el ámbito de la Justicia. Barcelona: Departament de Justicia, Generalitat de Catalunya.

Font-Barrot, Alfred; Pérez-Triviño, José Luis (2009). El Derecho para no Juristas: una guía para entender el sistema jurídico. Barcelona: Ediciones Deusto.

Forrest, Catherine B. (2021). When Machines Can Be Judge, Jury, and Executioner: Justice in the Age of Artificial Intelligence. Hackensack: World Scientific Publishing.

Gómez-Pérez, Asunción (2023). Inteligencia artificial y lengua española. Madrid: Safekat, S.L.

Habermas, Jürgen (2023). Una historia de la filosofía, vol.1. Madrid: Trotta.

Lefebvre (2024). Regulación de la IA y la protección de datos en España. Madrid: Lefebvre. https://lefebvre.es/tienda/libros-derecho-pdf-gratis/regulacion-de-la-ia-y-la-proteccion-de-datos-en-espana

Knuth, Donald (1997). The art of computing programming: fundamental algorithms. Reading: Addison-Wesley.

Real Decreto-ley 6/2023, de 19 de diciembre, por el que se aprueban medidas urgentes para la ejecución del Plan de Recuperación, Transformación y Resiliencia en materia de servicio público de justicia, función pública, régimen local y mecenazgo. (2023). // Boletín Oficial del Estado. 303 (20 de diciembre de 2023), 167808-167994.

Reglamento de Inteligencia Artificial (2024). Resolución legislativa del Parlamento Europeo, de 13 de marzo de 2024, sobre la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))

Robles, Gregorio. (2021). Teoría del Derecho. Fundamentos de teoría comunicacional del Derecho. Volumen III. Aranzadi: Pamplona.

Russell, Stuart, Norvig, Peter. (2021). Artificial intelligence: a modern approach. Hoboken: Pearson.

Saarenpää, Ahti. (2024). Some Difficulties in Reading the Law in the Age of Artificial Intelligence. // Schweig hofer, Erich; Eder, Stefan; Costantini, Federico; Schmaut zer, Felix; Pfister, Jonas (eds.). Sprachmodelle: Juristische Papageien oder mehr? Tagungsband des 27. Internationalen Rechtsinformatik Symposions. Editions Weblaw: Berna, 39-46.

Sanchis-Crespo, Carolina. (2023). Inteligencia artificial y decisiones judiciales: crónica de una transformación anunciada. // Scire. 29:2, 65-84.

Sandberg, Russell. (2023). A Historical Introduction to English Law.Genesis of the Common Law. Cambridge University Press: Cambridge.

Simó-Soler, Elisa; Rosso, Paolo. (2022). Inteligencia artificial y derecho: entre el mito y la realidad. La destrucción algorítmica de la humanidad. // Diario La Ley. 9982, 2022, 10 p.

Zekos, Georgios I. (2022). Advanced Artificial Intelligence and Robo-Justice. Cham: Springer.

Ziwei, Ji; Nayeon, Lee; Frieske, Rita; Tiezheng, Yu; Dan, Su; Yan, Xu; Etsuko, Ishii; Ye, Jin Bang; Madotto, Andrea; Fung, Pascale. (2023). Survey of Hallucination in Natural Language Generation. // ACM Comput. Surv. 55, 12, Article 248 (March 2023), 38 p.

Enviado: 2023-04-20. Segunda versión: 2024-05-31.
Aceptado: 2024-06-03.

Um panorama bibliométrico da proteção de dados e da privacidade em contexto de avanço da inteligência artificial

A bibliometric overview of data protection and privacy in the context of the advance of artificial intelligence

AIRES JOSÉ ROVER

Universidade Federal de Santa Catarina, Facultad de Derecho, aires.j.r@ufsc.br

Resumen

Durante el proceso de establecimiento de una nueva sociedad, surgen varios temas dignos de análisis y reflexión, destacando entre ellos el papel de la inteligencia artificial, especialmente el aprendizaje automático y sus aplicaciones predictivas. Aunque se ha explorado ampliamente la protección de datos personales y la privacidad en la literatura científica, existe una notable ausencia de estudios que conecten este análisis con las últimas técnicas de inteligencia artificial. En este sentido, proponemos llevar a cabo un mapeo más cuantitativo que cualitativo de las publicaciones científicas más influyentes que abordan estos temas interrelacionados. Para cumplir con este objetivo, hemos optado por utilizar un enfoque bibliométrico, llevado a cabo durante los meses de marzo y abril de 2024, empleando un método de análisis inductivo y adoptando el procedimiento de estudio de caso.

Palabras clave: Revisiones bibliográficas. Protección de datos. Privacidad. Inteligencia artificial.

Abstract

A new societal landscape is taking form, marked by a myriad of emerging themes warranting thorough examination. Chief among these discussions is the pivotal role of artificial intelligence (AI), particularly machine learning and its predictive applications. Despite extensive exploration of issues pertaining to personal data protection and privacy across scientific literature, there remains a conspicuous dearth of studies integrating this analysis with the latest AI methodologies. Hence, this study endeavors to undertake a predominantly quantitative mapping of the most influential scientific works addressing these interconnected themes. Our approach, conducted over the span of March and April 2024, employs a bibliometric methodology, prioritizing an inductive analysis framework and adopting the case study methodology to fulfill this objective.

Keywords: Bibliographic reviews. Data protection. Privacy. Artificial intelligence.

1. Introdução

Uma nova sociedade está sendo moldada, e nesse processo emergem diversos temas que merecem análise e reflexão. No centro dessas discussões está o papel da inteligência artificial, com destaque para o machine learning e suas aplicações preditivas. Embora as questões relacionadas à proteção de dados pessoais e privacidade tenham sido amplamente exploradas em várias publicações científicas, é evidente a falta de estudos que conectem essa análise com o uso das mais recentes técnicas de inteligência artificial.

Assim, este estudo propõe-se a realizar um mapeamento, mais quantitativo do que qualitativo, das publicações científicas mais influentes que abordam esses temas interligados. Para atingir esse propósito, optamos por adotar uma abordagem bibliométrica, conduzida ao longo dos meses de março e abril de 2024, utilizando um método de análise indutiva e adotando o procedimento de estudo de caso.

2. Progresso técnico, riscos e desafios na sociedade contemporânea

A técnica ou tecnologia não são novidades na trajetória da humanidade; elas acompanham o homem desde tempos imemoriais. A emergência da técnica decorre da limitação sensorial inerente ao ser humano. Carente de sentidos, o homem depende da adaptação inteligente do ambiente natural para superar suas deficiências (Gehlen, 1980). A tecnologia, definida como qualquer instrumento artificial que visa controlar a natureza em contraste com o mundo dos homens, é, portanto, uma construção cultural cujos objetos não são encontrados na natureza e têm por objetivo expandir os limites físicos e sensoriais do ser humano.

As demandas da sociedade moderna são extremamente exigentes, demandando a rápida elaboração de grandes quantidades de informações. Anteriormente, havia tempo para assimilar novas informações, permitindo uma aprendiza-

gem interna. No entanto, atualmente, esse processo de interiorização tornou-se inviável. Não é de se estranhar, portanto, que as relações entre as pessoas estejam se tornando cada vez mais superficiais. Tudo acontece em uma velocidade vertiginosa, e todos os processos sociais exigem um número crescente de decisões em intervalos de tempo cada vez mais curtos. A tecnologia, a economia e, por extensão, os demais sistemas sociais, refletem claramente essa revolução. Uma revolução caracterizada por um novo paradigma, composto por um conjunto de inovações técnicas, organizacionais e administrativas interrelacionadas, cujo fator-chave são os insumos baratos de informação decorrentes dos avanços em microeletrônica e telecomunicações, marcados pela redução dos custos relativos e pela disponibilidade universal (Castells, 1999, p. 54).

Esse processo de mediação tecnológica pode ser ainda mais radical, ultrapassando a visão clássica (prometéica) de domínio técnico da natureza, que mantém a fé no progresso material e na melhoria das condições humanas. Estamos potencialmente vivendo uma era fáustica da tecnociência, caracterizada por um impulso cego em direção ao domínio e à apropriação total da natureza, tanto externa quanto internamente ao corpo humano. Baseada em inteligência artificial e outras tecnologias disruptivas, a busca é pela transcendência do ser humano, uma verdadeira superação de suas limitações materiais, por meio da decifração do mistério da vida. Isso instaura uma forma de "biopoder", baseada na possibilidade de surgimento de "sociedades de controle" (Medeiros, 2003, p. 249).

Erros, falhas, riscos e perigos são inerentes a qualquer processo de transformação. Como Riobaldo, em "Grande Sertão: Veredas", afirmava, viver é perigoso. A inteligência humana permitiu a organização e a dominação por meio do trabalho, viabilizando o avanço da tecnologia. Esta se tornou um fator preponderante no processo de produção e transformação da humanidade, reduzindo os perigos, mas aumentando os riscos. O perigo é o risco que se concretiza. No entanto, a preocupação mais imediata é com a possível substituição ou domínio do ser humano por suas criações mecânicas. As máquinas certamente não substituirão o homem, mas o envolverão completamente, conferindo-lhe maior poder sobre a natureza e a sociedade. O verdadeiro risco reside nos processos que apenas as máquinas podem executar ou cujo controle humano é precário. A ameaça de falta de controle sempre estará presente. O que fazer então? Proibir simplesmente pesquisas que possam levar a essas situações? Ou arriscar até certo ponto e aprimorar os mecanismos de controle e vigilância?

A palavra-chave diante desses riscos é responsabilidade. Ela constitui o antídoto para transformar um risco em perigo. Quem são os agentes responsáveis pelas consequências de seus atos e omissões em diversos níveis? Definir esse cenário é uma tarefa regulatória complexa, uma vez que, cada vez mais, a responsabilidade das decisões recai sobre sistemas, e as pessoas tendem a se eximir dela. Não há mais ninguém para culpar em caso de falha: a culpa recai nos sistemas. Beck discute sobre uma sociedade que entra em uma fase de modernização reflexiva, tornando-se tema para si mesma e fonte de instabilidades e riscos provocados pelas novidades tecnológicas e organizacionais (2002, p. 21). Por exemplo, o princípio da precaução encontra seus limites nessa sociedade do risco, que demanda uma reflexão sobre si mesma. Assim, para ser contra o uso de determinada tecnologia, não é necessário possuir conhecimento, enquanto para ser a favor, é preciso possuir um entendimento profundo. O problema reside na polarização ideológica e na falta de conhecimento, o que dificulta a aplicação responsável desse princípio.

Portanto, é essencial aumentar a transparência na produção e distribuição de informações, facilitar a publicação de dados e proteger as informações de caráter privado. Essas são medidas de um regime aberto e de uma sociedade que se organiza de forma transparente e responsável. O avanço das tecnologias digitais pode impulsionar esse movimento. Como destacado por Rover (1995, p. 45), a humanidade há muito tempo almeja a utopia de um mundo universal, onde as pessoas possam estar mais conectadas sem perder sua autonomia, e onde o conhecimento, produto dessa autonomia, possa ser democratizado ao máximo.

Assim, o progresso técnico não é intrinsecamente bom nem mau, mas sim um instrumento cultural que, dependendo de seu uso, pode contribuir para o desenvolvimento humano em geral.

3. Privacidade e proteção de dados diante da inteligência artificial

A questão da privacidade e da proteção dos dados emerge como um tema central e controverso diante do avanço da inteligência artificial e suas técnicas. As lacunas na proteção legal do privado se aprofundam à medida que a capacidade de troca e difusão de informação se amplia. No entanto, é importante examinar os tipos de privacidade que a lei busca salvaguardar: (1) Informações sobre atos em geral que o indivíduo inevitavelmente pratica em público; (2) Informações sobre a vida pessoal que se deseja manter privativas. Enquanto a lei demonstra ser mais eficaz na

proteção do segundo tipo, o controle sobre o primeiro caso é menos eficiente, uma vez que as informações tornam-se públicas ao serem retiradas do âmbito privado, perdendo-se gradualmente o controle sobre elas.

Com o advento da Internet e, mais recentemente, da inteligência artificial em rede, as facilidades para obtenção de informação privada aumentam significativamente. Diversas são as formas de obter informações sobre os usuários, seja por meio de registros explícitos ou das pegadas deixadas durante o uso da rede. Embora as informações sejam geralmente fornecidas voluntariamente pelos usuários, é notório que muitos não estão preocupados em ocultar dados sobre suas vidas, entregando-os prazerosamente em troca de alguma vantagem social. Muitos justificam esse compartilhamento como forma de facilitar a navegação e receber sugestões personalizadas, orientadas por seus interesses.

No entanto, esse cenário suscita um paradoxo entre privacidade e liberdade de expressão. Enquanto é desejável evitar qualquer forma de censura na Internet, é essencial questionar quem determina o que é verdade ou relevante para o usuário. A liberdade de escolha entre diversas opções parece ser um direito fundamental, porém, há interesses políticos e econômicos em restringir essa liberdade, o que pode levar a um cenário reminiscente de um possível Big Brother.

Diante desses desafios, torna-se crucial promover mais liberdade e transparência para proteger os dados pessoais e a privacidade dos cidadãos, sem comprometer a resposta adequada aos desafios impostos pela intersecção entre tecnologia, privacidade e liberdade de expressão.

4. Inteligência Artificial apoiada por técnicas como aprendizado de máquina, Big Data e Large Language Models

A Inteligência Artificial é uma disciplina que recentemente alcançou a maturidade. Existem várias definições, sendo o paradigma da inteligência humana sua referência principal. Conforme Minsky (1985) afirmou, podemos defini-la como a ciência da construção de máquinas capazes de realizar tarefas que demandam inteligência, tal como seriam realizadas por seres humanos. Por outro lado, é também o campo de estudo que busca simular processos inteligentes ou de aprendizagem em máquinas, ou ainda tornar os computadores capazes de executar tarefas nas quais os seres humanos atualmente se destacam. Isso abrange habilidades como agir como especialistas, compreender e comunicar em linguagem natural, e reconhecer padrões como a escrita. Assim, há uma ampla gama de áreas de aplicação, ou melhor, desafios enfrentados por essa tecnologia: processamento de linguagem natural, reconhecimento de padrões (incluindo assinaturas, vozes e impressões digitais), robótica, execução de tarefas, resolução de problemas gerais ou especializados, bases de dados inteligentes, bancos de conhecimento, jogos, entre outros.

O ato de conhecer envolve três componentes: uma representação simbólica do objeto conhecido, uma inferência sobre ele e a capacidade de aprendizagem. Em termos de pesquisa, essa divisão é confirmada pelo foco em três grandes áreas dentro da inteligência artificial: representação do conhecimento, raciocínio e aprendizado (Rover, 2001, p. 108).

Os sistemas de inteligência artificial se valem da heurística, uma técnica utilizada para otimizar os processos de busca, ainda que em detrimento da perfeição ideal. Isso reflete o modo como os seres humanos interagem com o mundo. Essa abordagem serve como base para a implementação dos métodos dedutivo, indutivo e abdutivo. É um processo de compreensão do mundo que utiliza um conjunto definido de regras sobre um conhecimento específico. Assim, raciocinar implica em manipular informações (julgamentos, reconhecimentos), definir uma busca em um espaço de estados e inferir conclusões (Rover, 2001, p. 108).

Quanto ao aprendizado, em termos gerais, é a capacidade de um agente ou sistema melhorar seu desempenho (D) em uma classe de tarefas (T) como resultado da experiência (P). Existem diversas técnicas para implementar o aprendizado nos sistemas. Elas visam melhorar o desempenho, aumentar a robustez e eficiência dos sistemas, aprendendo novas regras e gerando novas soluções (Rover, 2001, p. 63).

Por fim, a representação do conhecimento é crucial. O conhecimento precisa ser representado dentro da máquina para que possa ser processado e apresentar as conclusões desejadas. Isso envolve escolhas de modelagem ontológicas (fonte, alcance, orientação, nível, resolução), de comportamento (precisão, incerteza) e principalmente de representação (equações, associações, procedimentos). Existem várias técnicas de representação, como sistemas de produção, redes semânticas, quadros (frames) e lógica (Rover, 2001, p. 63).

Há muitas técnicas para implementar sistemas inteligentes, e geralmente são introduzidas inovações às técnicas tradicionais ou sistemas híbridos que combinam várias delas. Entre as mais discutidas estão os Sistemas Baseados em Regras, os Sistemas Baseados em Casos e as Re-

des Neurais. Cada técnica tem sua aplicação específica, e é importante evitar a tentativa de substituição arbitrária entre elas, pois suas características são distintas.

A técnica de aprendizado de máquina, tão discutida e testada nos últimos anos, não é uma novidade em si, mas sim uma evolução de algoritmos antigos em sistemas relacionados, como redes neurais ou algoritmos genéticos. O que é novo é a capacidade de acessar e interpretar dados massivos, estruturados e não estruturados, e utilizá-los para fazer predições automáticas (análise preditiva) sem necessidade de nova programação (Assunção, 2018, p. 23). O aprendizado de máquina geralmente é supervisionado por humanos, que preparam e rotulam os dados antes de treinar o algoritmo. Esse processo envolve ajustes com base nos resultados, identificando gradualmente o melhor caminho para alcançar um objetivo específico. A evolução dos algoritmos e o acesso facilitado a grandes bases de dados estão permitindo a automação de decisões que antes eram exclusivamente humanas. É fundamental garantir a correção dos resultados dessas aplicações, utilizando dados imparciais e corrigindo os algoritmos conforme necessário.

As máquinas de aprendizado, embora conhecidas há muito tempo, só recentemente puderam ser implementadas e utilizadas com eficiência devido à disponibilidade de grandes bases de dados para uso geral. Neste contexto, o termo Big Data refere-se à gestão de dados em larga escala, com múltiplos conteúdos e produção em alta velocidade (Ribeiro, 2014, p. 101). A análise desses dados visa à predição de fenômenos com base em correlações diretas, sem necessidade de amostragens menores. No contexto do governo eletrônico, embora o termo Big Data não seja comum, a ideia de "governo aberto" se aproxima, pois os dados governamentais são frequentemente considerados Big Data, e as características mencionadas anteriormente também se aplicam. No entanto, ao contrário do modelo tradicional de dados abertos do governo, nos quais a ênfase está na abertura pelos próprios órgãos governamentais, a abertura possibilitada pela tecnologia de Big Data é mais automatizada, permitindo acesso a um conhecimento que antes era inacessível devido à sua complexidade.

Além das máquinas de aprendizado, uma vertente tecnológica que vem ganhando destaque são os LLMs (Large Language Models), modelos de linguagem de grande escala que utilizam técnicas avançadas de processamento de linguagem natural para entender e gerar texto de forma cada vez mais próxima à humana. Esses modelos, como o GPT (Generative Pre-trained Transformer), têm sido aplicados em uma variedade de campos, desde assistentes virtuais até tradução automática e geração de texto criativo. Com suas capacidades de processamento e compreensão de grandes volumes de dados textuais, os LLMs representam uma ferramenta poderosa para análise e geração de insights em meio ao Big Data.

Mais precisamente, os LLM (Large Language Models), são modelos de aprendizado de máquina treinados em vastos conjuntos de dados. Eles são utilizados para gerar linguagem para interações com humanos e para desenvolver contexto, possibilitando respostas rápidas em plataformas de IA generativa. Tecnologias como ChatGPT, Gemini, Copilot, DALL-E e Midjourney dependem desses LLMs para operar. Essas redes neurais adquirem conhecimento ao longo do tempo e produzem respostas em texto, imagem, vídeo e até mesmo em código de programação. Por outro lado, o ChatGPT é um sistema de IA gratuito que permite conversas envolventes, oferece insights e automatiza tarefas. Ele se baseia em um LLM com 175 bilhões de parâmetros, treinado em uma extensa base de dados, e é capaz de gerar textos sofisticados e aparentemente inteligentes.

Dessa forma, o ChatGPT é capaz de conduzir conversas cada vez mais naturais e sofisticadas com os usuários, adaptando-se ao contexto e respondendo a uma ampla gama de perguntas e solicitações. Sua capacidade de compreender a linguagem humana e gerar respostas relevantes o torna uma ferramenta valiosa não apenas para entretenimento, mas também para suporte ao cliente, educação e até mesmo assistência em tarefas complexas. Assim, o ChatGPT exemplifica como os LLMs estão transformando a maneira como interagimos com a tecnologia e exploramos o vasto universo de dados disponíveis.

No entanto, junto com os avanços e benefícios trazidos pelas máquinas de aprendizado e os LLMs, também surgem preocupações e riscos significativos. Um dos principais receios está relacionado à privacidade e segurança dos dados, especialmente em um contexto de Big Data, onde enormes quantidades de informações pessoais podem ser coletadas, armazenadas e analisadas sem o consentimento adequado dos usuários. Além disso, há preocupações éticas sobre o uso dessas tecnologias para manipulação de opiniões, disseminação de desinformação e até mesmo criação de conteúdo prejudicial, como deepfakes. Outro risco é a amplificação de preconceitos e discriminação, uma vez que os modelos de aprendizado de máquina podem reproduzir e até mesmo agravar viés presentes nos dados de treinamento. Portanto, é crucial que os desenvolvedores e usuários estejam atentos a

essas questões e implementem medidas adequadas para mitigar esses riscos enquanto aproveitam os benefícios oferecidos pela tecnologia.

Enfim, as soluções inteligentes que avançam e a enorme quantidade de dados disponíveis devem ser utilizadas por óbvio, não apenas como uma oportunidade, mas como uma necessidade. No entanto, é crucial avançar nessa direção de forma técnica e ética, a fim de evitar abusos e garantir a defesa dos direitos individuais. Por isso, já temos algumas iniciativas de regulamentação para garantir sua utilização responsável e a preservação dos direitos individuais (SARLET, 2022, p. 25). Por exemplo, no Brasil, várias medidas estão em curso nesse sentido. A Lei Geral de Proteção de Dados (LGPD), em vigor desde 2020, estabelece princípios para o tratamento de dados pessoais, cuja aplicação se estende também ao uso de dados na IA, apesar de não ser seu foco principal. A Autoridade Nacional de Proteção de Dados (ANPD) é a entidade encarregada de implementar a LGPD e promover uma cultura de proteção de dados no país, acompanhando de perto o debate sobre a regulamentação da IA. O Projeto de Lei nº 21/2020, já aprovado na Câmara dos Deputados e aguardando avaliação no Senado Federal, estabelece um marco legal para o desenvolvimento e uso da IA no Brasil, delineando princípios, direitos e deveres, enquanto o Projeto de Lei 2338/2023, em tramitação, busca estabelecer normas específicas para a IA, com foco em princípios como respeito à dignidade humana e proteção da privacidade. A Política Nacional de Inteligência Artificial (PNIA), lançada pelo governo brasileiro em 2023, define diretrizes para o desenvolvimento e uso da IA no país, reconhecendo a importância da proteção de dados e da ética. Na área da saúde, a regulação da IA demanda transparência na coleta de dados, gestão de riscos, validação externa de dados e proteção da privacidade.

No cenário global, diversas iniciativas também merecem destaque. O Regulamento Geral de Proteção de Dados (RGPD), em vigor na União Europeia desde 2018, é um marco abrangente na proteção de dados, incluindo disposições específicas para o tratamento de dados na IA. A Lei de Proteção de Dados Pessoais e Privacidade da Califórnia (CCPA), em vigor desde 2020, concede direitos aos consumidores californianos em relação aos seus dados pessoais, com disposições específicas para o uso de dados na IA. Além disso, diversas outras nações e organizações internacionais estão empenhadas em desenvolver iniciativas para regular a IA e proteger os dados, como as diretrizes da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) e os princípios da UNESCO sobre IA. Em suma, várias nações estão adotando estratégias para promover o desenvolvimento da IA, envolvendo governos, indústrias e universidades.

5. Metodologia da bibliometria utilizada

A bibliometria é uma técnica essencial para medir índices que indicam a produção e disseminação do conhecimento científico (Fonseca, 1986, p. 73). Estes índices permitem a análise detalhada de um campo científico específico, revelando características como o crescimento temporal da produção científica, a produtividade de autores e instituições, a colaboração entre pesquisadores e instituições, o impacto das publicações, bem como a análise e avaliação das fontes de disseminação de trabalhos e a distribuição da produção científica em diversas fontes e temáticas. Esta análise pode revelar a evolução e tendências nesse campo (Bufrem & Prates, 2005, p. 12).

Para explorar esse tema, realizamos uma análise através do levantamento de artigos indexados no Google Acadêmico, uma base reconhecida pela sua amplitude e pela variedade de revistas que indexa. As buscas foram restritas aos títulos das publicações, utilizando os operadores OR e AND, Essas escolhas restringem os resultados, porém dão maior precisão aos mesmos.

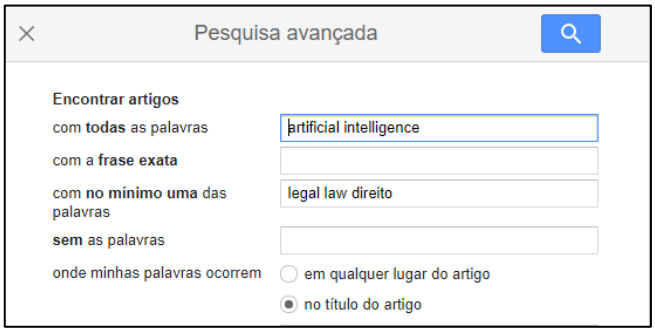

Figura 1. Janela da pesquisa avançada do Google Acadêmico

Dada a diversidade de palavras-chave, estas foram utilizadas de forma estratégica para organizar os resultados. As buscas foram conduzidas com dois polos: um principal, formado por termos equivalentes ligados pelo operador OR, e um secundário, composto por termos não equivalentes, exigindo uma busca distinta para cada um. A inclusão de termos no plural ampliou o espectro de estudos pertinentes ao tema.

Durante a análise dos resultados, não impusemos restrições temporais aos artigos. Quando os resultados eram escassos, repetições foram desconsideradas para evitar distorções.

As palavras-chave selecionadas abordam tanto os temas relativos à inteligência artificial quanto os relativos à proteção de dados pessoais.

6. Discussão dos resultados

A busca realizada (B1) teve como objetivo fornecer um panorama amplo da interação entre o Direito e os temas de inteligência artificial, Large language models e chatgpt. Para isso, as palavras-chave foram categorizadas em dois grupos: os termos principais que abrangem o Direito de forma geral (legal e law) e os termos secundários relacionados à inteligência artificial. As buscas foram conduzidas para cada termo secundário, combinando-os com os termos principais usando o operador AND. Foram utilizados exclusivamente termos em inglés.

legal OR law	
AND	*Total de artigos*
artificial intelligence	2300
Large language models	81
chatgpt	88

Figura 2. Busca B1

Os resultados quantitativos revelam que houve um grande número de retornos com o termo artificial intelligence, com os seis primeiros artigos apresentando mais de 100 citações cada, e os demais mantendo uma quantidade próxima. Esse achado indica um alto interesse na interseção dessas duas áreas. No entanto, é importante notar que esses resultados são relevantes apenas em termos de contexto, uma vez que os artigos mais citados apresentam uma visão bastante geral da relação entre as duas temáticas. Seria preciso uma análise dos demais artigos para identificar se especificam mais as temáticas.

Por outro lado, a análise do termo Large language models em relação ao Direito resultou em uma diminuição significativa nos retornos, mesmo assim um número alto de artigos. Além disso, o número de citações desses artigos é até razoável, entre 10 e 30. Isso sugere que a temática é mais específica e, portanto, há um interesse relativo nesse campo de estudo, embora não seja negligenciável. A maioria dos artigos exploram o papel e o potencial dos grandes modelos de linguagem na esfera jurídica, abordando diversos aspectos que vão desde a compreensão do conhecimento jurídico até a aplicação prática desses modelos. O primeiro texto discute o impacto do tamanho e do método de treinamento dos modelos de linguagem na sua performance, buscando identificar os aspectos mais influentes nesse desempenho. Em seguida, um outro estudo apresenta o ChatLaw, um modelo de linguagem jurídica desenvolvido especialmente para o contexto legal chinês, destacando a importância da qualidade dos dados para sua eficácia. Outro trabalho propõe o Legalbench, um benchmark colaborativo para medir o raciocínio jurídico em grandes modelos de linguagem, ressaltando como avanços nessa área estão levando profissionais do direito a reconsiderar suas práticas. Além disso, há uma análise sobre a integração dos grandes modelos de linguagem no campo jurídico, destacando seus resultados promissores em tarefas como análise de documentos legais e revisão de contratos. Um estudo mais específico examina as "ficções legais", ou seja, distorções na interpretação de fatos legais por parte desses modelos, enquanto outro destaca a emergência de capacidades de compreensão jurídica em evolução nos modelos de linguagem. Também são abordadas propostas para ensinar esses modelos a prever julgamentos legais, bem como desafios enfrentados por eles e métodos de avaliação. Finalmente, um estudo chinês apresenta um benchmark de modelos de linguagem jurídica, mostrando diferentes focos e aplicações dentro desse contexto. Esses trabalhos refletem um interesse crescente e multidisciplinar na interseção entre inteligência artificial e direito, explorando tanto os desafios quanto as oportunidades que essas tecnologias representam para a prática jurídica contemporânea.

Já a análise do termo chatgpt resultou também uma diminuição nos retornos, na mesma faixa do termo anterior. Além disso, o número de citações desses artigos é parecido com o caso anterior, ocorrendo entre 10 e 40 vezes. Da mesma forma acima, os artigos abordaram várias facetas da integração do ChatGPT no campo jurídico, revelando tanto as oportunidades quanto os desafios associados a essa tecnologia. Primeiramente, destaca-se a capacidade do ChatGPT de auxiliar professores de direito em tarefas comuns, demonstrando um desempenho positivo em prompts relacionados ao serviço, o que sugere que a ferramenta pode oferecer suporte próximo aos professores de direito. Em contrapartida, há uma discussão sobre os desafios legais e éticos decorrentes do uso de modelos de linguagem grandes, como o ChatGPT, incluindo preocupações com parrots estocásticos e alucinações, que podem impactar questões legais e sociais de maneira significativa. Outro artigo explora como o ChatGPT pode ser aplicado na educação jurídica e na prática, oferecendo exemplos de prompts e conselhos sobre seu uso. Também são discutidas implicações éticas e legais específicas do uso do ChatGPT na urologia, destacando a necessidade de considerar cuidadosamente os impactos éticos e legais ao integrar essa tecnologia em áreas

sensíveis. A pesquisa também aborda o papel do ChatGPT no campo jurídico, destacando suas possíveis utilidades para paralegais e assistentes jurídicos em diversas tarefas legais. Além disso, são examinadas as implicações do ChatGPT para a educação e prática jurídica, considerando como essa tecnologia pode influenciar o ensino e o exercício do direito. A discussão sobre inteligência artificial no contexto jurídico também levanta questões sobre a tomada de decisões legais e a autenticidade do ChatGPT em investigações forenses, evidenciando a necessidade de uma avaliação cuidadosa de sua aplicação. Por fim, são exploradas as implicações do ChatGPT para os serviços legais e a sociedade, destacando seu potencial como ferramenta valiosa, mas também ressaltando a importância de considerar questões éticas e legais ao implementá-lo.

Inteligência artificial	
AND	*Total de artigos*
proteção de dados	10
privacidade	41

Figura 3. Busca B2

Em seguida fizemos buscas mais focadas. A busca B2 relacionou o termo principal inteligência artificial com os secundários proteção de dados e privacidade, todos os termos em português.

O termo, proteção de dados, resultou 10 artigos, mostrando ser pequena a amostragem e com poucas citacoes. Quanto a análise dos titulos dos artigos, ficou evidente a interconexão entre os conceitos de Direitos Humanos, inteligência artificial e privacidade em diversos contextos. Os artigos abordam temas que vão desde os riscos e implicações da inteligência artificial e dos algoritmos para a privacidade, até questões específicas como a utilização da inteligência artificial no âmbito da saúde e seus limites em relação à privacidade dos pacientes. Além disso, há discussões sobre a relativização da privacidade em situações como a violência doméstica, onde se questiona a ponderação entre a garantia à integridade física e a preservação da privacidade dos envolvidos. A aplicação da inteligência artificial nas redes sociais também é explorada em relação à proteção da privacidade e dos dados pessoais dos usuários, destacando a importância de políticas de privacidade transparentes. Outros temas incluem os desafios enfrentados pela legislação de proteção de dados na era da inteligência artificial, especialmente em relação às violações de privacidade decorrentes do uso de robôs e das chamadas telefônicas automatizadas. A discussão sobre ética, transparência e responsabilidade no uso da inteligência artificial também é abordada em relação ao equilíbrio entre eficiência e privacidade na luta contra as fake news.

Já o termo privacidade, resultou 41 amostras, com 3 deles tendo algumas citacoes. Alguns estudos destacam o papel das leis gerais de proteção de dados como possíveis vetores para a regulamentação da inteligência artificial, investigando se o princípio da precaução, aliado à accountability e aos relatórios de impacto à proteção de dados, poderia ser o portal de entrada para essa regulamentação. Outros artigos focam na aplicação da Lei Geral de Proteção de Dados Pessoais como ponto de partida para a regulação da inteligência artificial em setores específicos, como a saúde, apontando para a necessidade de desenvolvimento de tecnologias e legislações que garantam a salvaguarda dos direitos individuais. Além disso, há estudos que exploram a relação entre a inteligência artificial, a proteção de dados pessoais e a responsabilidade na era digital, destacando a importância de conciliar o impacto da inteligência artificial sobre as pessoas com o respeito aos direitos fundamentais estabelecidos pela Constituição Federal. Ainda, há reflexões sobre os desafios suscitados pelo uso da inteligência artificial e do big data no contexto da COVID-19, especialmente no que diz respeito à proteção adequada dos dados pessoais e à alocação de responsabilidade por eventuais danos. Esses estudos também abordam a mudança de paradigma na proteção de dados e o uso da inteligência artificial a partir de um modelo constitucional, visando garantir a proteção dos direitos fundamentais envolvidos. Ademais, é discutido o papel da accountability e do direito fundamental à proteção de dados pessoais como limites ao uso da inteligência artificial na relação de emprego, enfatizando a necessidade de mecanismos que garantam a transparência e a prestação de contas no tratamento automatizado de dados.

Artificial intelligence	
AND	*Total de artigos*
data protection	144
privacy	354

Figura 4. Busca B3

Na busca B3, foram utilizados exclusivamente termos em inglês, sendo o termo principal "artificial intelligence" e os secundários "data protection" e "privacy".

O termo "data protection" resultou em 144 artigos, sendo os primeiros 10 muito citados (de 20

a 50 citações). Estes 10 artigos destacam várias questões, como a crescente importância da regulação da inteligência artificial (IA), especialmente no contexto da proteção de dados pessoais. Um dos debates examina a relação entre o Regulamento Geral de Proteção de Dados (GDPR) e a IA, evidenciando preocupações sobre a capacidade do GDPR em lidar eficazmente com os avanços rápidos e significativos na IA. Outra discussão sobre a ética na governança de TI e a aplicação de modelos de governança legal, como o GDPR, revela a necessidade de uma abordagem abrangente que leve em consideração as especificidades da IA e da proteção de dados. Além disso, há uma análise sobre como a IA e o big data apresentam novos desafios para a proteção de dados, uma vez que permitem previsões sobre terceiros com base em dados anônimos. Esses artigos destacam a urgência de uma abordagem multidisciplinar e colaborativa para lidar com as interseções entre IA, proteção de dados e cibersegurança, a fim de garantir um quadro legal eficaz e ético.

Já o termo "privacy" resultou em 354 amostras, sendo os 10 primeiros muito citados (de 100 a 300 citações). É notável que este termo desperta grande interesse na interseção entre privacidade e inteligência artificial (IA). Os artigos abordam diversos contextos, incluindo saúde, destacando técnicas e aplicações para preservar a privacidade dos dados. Há também discussões sobre privacidade e IA, enfatizando a importância de garantir a segurança dos dados em um cenário de crescente uso de algoritmos de IA. Uma abordagem específica sobre privacidade na área da saúde ressalta os desafios adicionais enfrentados ao lidar com informações sensíveis dos pacientes. Ao explorar os riscos associados à IA para a privacidade, os textos apontam como a prática da IA pode afetar diretamente a segurança dos dados pessoais, levantando preocupações sobre potenciais violações e uso inadequado das informações. Além disso, discute-se os riscos para a privacidade e a democracia, destacando as implicações mais amplas do uso da IA na manipulação de informações e influência em processos democráticos. Por fim, ao mencionar a aplicação de técnicas como a privacidade diferencial em IA, os textos sugerem abordagens para mitigar os riscos à privacidade, evidenciando a necessidade contínua de inovação e pesquisa nesse campo para equilibrar os avanços tecnológicos e a proteção dos direitos individuais.

A busca B4 explorou a interconexão entre o termo central "machine learning" e os conceitos secundários de "data protection" e "privacy", ambos discutidos anteriormente. Surpreendentemente, apesar da natureza altamente específica dessas áreas, os resultados foram extensos. Isso ressalta a distinção marcante entre "machine learning" e "inteligência artificial", onde o primeiro é notavelmente mais preciso, uma distinção que tem implicações significativas.

machine learning	
AND	*Total de artigos*
data protection	68
privacy	1130

Figura 5. Busca B4

Começando com "data protection", a busca resultou em 68 artigos. Os principais temas abordados nesses artigos variam desde a implementação do Regulamento Geral de Proteção de Dados (GDPR) de 2016 até questões cruciais sobre conformidade legal sem prejudicar a precisão das análises, especialmente em setores sensíveis como a pesquisa médica. Outros temas incluem a Privacidade por Design como uma abordagem essencial para equilibrar insights úteis e proteção de dados, além do potencial do aprendizado de máquina federado para compartilhamento de dados entre instituições, respeitando os requisitos de privacidade. No entanto, destaca-se a lacuna entre a regulamentação de proteção de dados e a prática dos algoritmos de aprendizado de máquina, apontando para a necessidade de uma definição mais precisa de dados e informações para uma regulamentação mais eficaz. Também é levantada a questão da soberania cognitiva e a crescente preocupação com o rastreamento e perfilamento de indivíduos.

Quanto ao termo "privacy", os resultados foram ainda mais impressionantes, com 1130 amostras identificadas. A alta quantidade de citações desses artigos destaca a forte ligação entre o aprendizado de máquina e a privacidade, evidenciando a relevância dessa discussão. Os temas abordados incluem a integração de técnicas de privacidade no aprendizado de máquina, uma revisão da literatura existente sobre privacidade nesse domínio, ataques à privacidade em modelos de aprendizado de máquina, propostas de serviços que preservam a privacidade dos dados dos usuários, segurança e privacidade no contexto do aprendizado de máquina, e ameaças potenciais à privacidade, soluções propostas e desafios futuros nesse campo em constante evolução.

É notável que os aspectos puramente jurídicos e regulatórios não sejam predominantes nesses resultados, sugerindo que a ênfase está na

adoção de práticas estabelecidas e reconhecidas. No entanto, a confirmação dessa hipótese requer uma análise completa dos artigos em estudos futuros.

large language model OR chatgpt	
AND	Total de artigos
data protection	7
privacy	81

Figura 6. Busca B5

A análise B5 examinou a relação entre os termos "large language models" ou ChatGPT, considerados sinônimos nesta discussão, e os termos "data protection" e "privacy", abordados separadamente. Dada a distinção jurídica entre os termos, as amostras obtidas foram notavelmente distintas.

Em relação a "data protection", apenas sete artigos foram encontrados, com poucas citações, exceto pelo primeiro, que obteve 20 citações. Estes artigos exploram várias perspectivas, começando com a proteção de dados em chatbots baseados em inteligência artificial, especialmente o ChatGPT da OpenAI. Outras preocupações incluem o impacto dos grandes modelos de linguagem, como o ChatGPT, na segurança dos dados dos usuários, desafios e oportunidades para legisladores alemães na balança entre proteção de dados e uso do ChatGPT, e o impacto do ChatGPT nas leis de privacidade e proteção de dados, destacando questões específicas para garantir conformidade e melhorias na privacidade. Além disso, avalia-se se o ChatGPT está influenciando efetivamente as preocupações com segurança cibernética e proteção de dados em diferentes regiões, como a União Europeia, os EUA e a China. Também são explorados os princípios de proteção de dados, especialmente em relação ao GDPR, com os quais o ChatGPT está em conformidade. Por fim, discute-se técnicas de marca d'água em dados de texto em grandes modelos de linguagem para proteger direitos autorais dos conjuntos de dados e garantir a privacidade dos usuários.

Quanto ao termo "privacy", os resultados foram mais substanciais, com 81 amostras identificadas. Os cinco primeiros artigos foram citados entre 40 e 160 vezes, enquanto os demais tiveram até 10 citações. Os primeiros artigos abordam diversas questões relacionadas ao uso do ChatGPT, com menção limitada às LLMs. O primeiro tópico discute sustentabilidade, avaliando o impacto ambiental e a viabilidade a longo prazo dessas tecnologias. Em seguida, a privacidade é destacada, com foco nas ameaças à privacidade decorrentes do uso desses sistemas, especialmente em relação à coleta e uso de dados pessoais. A divisão digital é mencionada como um fator que pode agravar disparidades sociais e econômicas. Por fim, questões éticas são levantadas, ressaltando a importância de considerações morais ao desenvolver e implantar sistemas de IA, especialmente aqueles que afetam diretamente a interação humana. Esses temas destacam a necessidade urgente de avaliação contínua e aprimoramento em todas as áreas mencionadas para garantir que o progresso tecnológico seja equitativo, sustentável, respeite a privacidade e seja ético.

É relevante observar que os aspectos puramente jurídicos e regulatórios são abordados de maneira genérica, com foco nos aspectos técnicos.

7. Considerações finais

A presente pesquisa teve como objetivo realizar um mapeamento quantitativo das publicações científicas sobre inteligência artificial, machine learning e large language model (LLM) em relação à privacidade e proteção de dados. Utilizou-se uma abordagem bibliométrica, priorizando buscas nos títulos das publicações e cruzando termos de maneiras diversas para avaliar a frequência dessas relações e seu significado qualitativo. Alguns resultados foram qualitativamente apresentados.

Os achados revelaram um avanço significativo na discussão científica desses temas técnicos em conexão com questões regulatórias e normativas. As publicações em língua inglesa foram o foco principal, com uma busca limitada em termos em português, resultando em um retorno ainda restrito.

Na relação entre a inteligência artificial e o direito os resultados demonstram que são temas já consolidados, destacando-se a análise dos desafios enfrentados pela legislação, especialmente diante do rápido avanço tecnológico, a relevância da Privacidade por Design e a importância de abordagens multidisciplinares para lidar com questões de cibersegurança e proteção de dados.

Já nos artigos relacionados ao aprendizado de máquina e sua relação com a proteção de dados e privacidade, embora mais restritos, também obtiveram bons resultados. Um pouco menos a relaçãco com os "Large Language Models" e o "ChatGPT", indicando um espaço importante de discussão ainda em aberto. Emergem aqui preocupações como alucinações e a necessidade de assegurar transparência e responsabilidade no

emprego dessas tecnologias. Preocupações éticas e legais são ainda bem gerais e não são predominantes, sugerindo que a ênfase está na discussão de temas mais práticos e técnicos.

Como é obvio, nossa conclusão não é geral e diz respeito apenas aos artigos mais citados. Em estudos futuros seria possível avançar para uma análise mais completa de todos artigos recuperados. Espera-se que os resultados deste estudo contribuam para a disseminação e mapeamento desses temas, apontando para um amplo campo de novas e futuras pesquisas nessa área interdisciplinar.

Referencias

Assunção, Luís. Machine learning, big data e inteligência artificial. I// Lex Machinae. https://www.lexmachinae.com/2017/12/08/machine-learning-big-data-e-inteligencia-artificial-qual-o-be.

Beck, Ulrich; Zolo, Danilo (2022). A sociedade global do risco. // Prim@ Facie. 1: 1, 18-29.

Bufrem, L.; Prates, Y. (2005). O saber científico registrado e as práticas de mensuração da informação. // Ciência da Informação. 34:2, 9-25, 2005.

Castells, Manuel (1999). A sociedade em rede. São Paulo: Paz e Terra.

Fonseca, E. N. (1986). Bibliometria: teoria e prática. São Paulo: Cultrix, Ed. da USP.

Gehlen, Arnold (1980). Man in the Age of Technology. New York: Columbia Univ. Press.

Medeiros, Nelma (2003). O Homem Pós-Orgânico: Quarta Ferida Narcísica? // Novamente Revista. 4/5, 247-252.

Minsky, M. (1985). The Society of Mind. New York, USA: Touchstone.

Ribeiro, Claudio José Silva (2014). Big Data: os novos desafios para o profissional da informação. // Informação & Tecnologia (ITEC). 1:1, 96-105.

Rover, Aires J. (2001). Informática no direito: inteligência artificial: introdução aos sistemas especialistas legais. Juruá Editor.

Rover, Aires José; Melo, Marco A M Ferreira de (1995). Perspectivas do uso da Internet no curso de direito. In: Revista seqüência.30, 65-79.

Sarlet, Ingo Wolfgang (2022). Inteligência Artificial, Proteção de Dados Pessoais e Responsabilidade na Era Digital. Série Direito, Tecnologia, Inovação e Proteção de Dados num Mundo em Transformação. eBook Kindle, 2022.

Enviado: 2024-05-06. Segunda versión: 2024-05-23.
Aceptado: 2024-05-23.

Análise do discurso pecheuxtiana: uma proposta metodológica na área da ciencia da informação

Pecheuxtian discourse analysis: a methodological proposal in the area of information science

Edina Rodrigues Lima (1), Daniel Martínez-Ávila (2), Blanca Rodríguez-Bravo (2), Olga Myllena Diniz Botelho Santana (3)

(1) Biblioteca da Presidência da República - Brasil, Praça dos Três Poderes - Palácio do Planalto - Anexo I Ala B Superior, Esplanada dos Ministérios, 70000000 - Brasília, DF - Brasil, edina.lima@presidencia.gov.br. (2) Área de Biblioteconomía y Documentación, Facultad de Filosofía y Letras, Universidad de León, Campus de Vegazana, s/n, 24071 León, España, {dmarta | blanca.rodriguez}@unileon.es. (3) PPGCI UFRJ/IBICT, Rua Lauro Müller, 455, Botafogo, Rio de Janeiro - RJ, 22290-160, Brasil, myllena.diniz@gmail.com

Resumen

En poco más de seis décadas, el Análisis del Discurso (AD) ha realizado una importante aportación teórica y metodológica a la investigación en Humanidades, Ciencias Sociales e Información. Basado en una comprensión más profunda del lenguaje, no como un sistema independiente, sino como un medio de producción y difusión de significados, dotado de material simbólico e ideológico, el AD permite entender la lingüística como un agente activo, resultado de la convergencia entre discurso, sujeto e historicidad. Desde esta perspectiva, este estudio presenta una propuesta para el tratamiento de la información, especialmente en el campo de la indización, de los objetos informativos a partir de la teorización de Pêcheux sobre la significación discursiva y la formación de la "fuerza imaginaria".

Palabras clave: Análisis del discurso. Análisis materialista del discurso. Ciencia de la información. Lenguaje. Lingüística. Discurso. Pêcheux, Michel.

Abstract

In just over six decades, Discourse Analysis (DA) has made a relevant theoretical-methodological contribution to research in the areas of Human, Social, and Information Sciences. By delving deeper into language, not as an independent system, but as a means of producing and disseminating meanings, endowed with symbolic and ideological material, DA enables the understanding of linguistics as an active agent, the result of the convergence between discourse, subject, and historicity. From this perspective, this study presents a proposal for subject representation, especially in the field of indexing, of informational objects based on the theorization of discursive meaning and the formation of Pêcheux's "imaginary force".

Keywords: Discourse analysis. Materialist discourse analysis. Information Science. Language. Linguistics. Discourse. Pêcheux, Michel.

1. Introdução

Ao considerar o caráter polissêmico da Análise do Discurso (AD), compreende-se a possibilidade de analisar os discursos a partir de diferentes conceitos e formulações teóricas, de modo a superar o enfoque dado ao uso da linguagem em si para um aprofundamento sobre os efeitos de sentido produzidos entre interlocutores. De tal forma, este estudo parte da concepção francesa da Análise Materialista do Discurso, concebida em meados da década de 60, sob a égide do filósofo Michel Pêcheux (1938-1983), como ferramenta a serviço de propostas teórico-metodológicas das mais diversas áreas, sobretudo, da Ciência da Informação. Nesta perspectiva, a linguagem é analisada a partir de sua materialidade discursiva, ideológica.

A AD joga luz sobre a construção do conhecimento e da sociedade, por meio da linguagem – não pela estrutura única da fala ou da escrita, mas pela convergência entre discurso, sujeito e história. Portanto, daquilo que se materializa por meio de sentidos e efeitos de sentidos – históricos e sociais (teoria marxista), realizados por sujeitos (teoria freudiana) e realizáveis por meio da materialidade da linguagem (teoria saussuriana) (Mazzola, 2009).

Não à toa, a emergência dessa corrente, dentro da Linguística, ocorre na segunda metade no século XX – em maio de 1968, na França, diante de manifestações contra a rigidez do sistema educacional e a favor de melhorias trabalhistas, sendo marcada por convulsões nas ruas e na academia –, quando linguistas rompem com a preocupação restrita à análise do interior do enunciado e passam a focar na enunciação discursiva. Ou seja, naquele momento, havia a passagem de uma "Linguística da frase" para uma "Linguística do discurso", superando, assim, a dicotomia entre

língua e fala – como sistema social e algo particular, respectivamente –, estabelecida por Saussure (Mazzola, 2009), para considerar os impactos das variáveis socioculturais sobre os interlocutores.

Desta forma, a AD compreende a língua no campo simbólico, enquanto sentido, para além da linguagem. Portanto, "o discurso implica em uma exterioridade à língua, encontra-se no social e envolve questões de natureza não estritamente linguística" (Fernandes, 2005, p. 12), de modo que, ao observarmos indivíduos antagônicos, em situação de debate, identificaremos que "as posições em contraste revelam lugares socioideológicos assumidos pelos sujeitos envolvidos, e a linguagem é a forma material de expressão desses lugares (Fernandes, 2005, p. 12). Assim, o escopo da AD "é determinado pelo campo dos espaços discursivos não estabilizados logicamente, dependendo dos domínios filosófico, sociohistórico, político ou estético, e, portanto, também dos múltiplos registros do cotidiano não estabilizado" (Pêcheux, 1998, p. 54).

Dito isso, neste trabalho, busca-se aprofundar as discussões e a compreensão sobre os elementos teóricos e metodológicos pelos quais é possível analisar a língua a partir de bases materiais, pautada na relação entre ideologia, sentido e história, para a construção dos discursos. Portanto, dedicada, muito mais, à produção discursiva do que ao conteúdo se trata de uma proposta relevante para os estudos na área da Ciência da Informação e outros campos afins da Documentação e a Comunicação como o trabalho em bibliotecas (incluindo especializadas como por exemplo de informação jurídica) ou no jornalismo.

2. A linguagem (mensagem) e seu funcionamento

2.1. O que é a linguagem

A linguagem é um sistema por meio do qual os homens manifestam seus sentimentos, ideias e vontade, seja pela fala, pela escrita ou por signos convencionados. Ao depender do contexto social em que a linguagem será dinamizada, o interlocutor poder-se-á manifestar por meio de uma linguagem formal, a qual exige um padrão normativo culto. A linguagem formal é utilizada como a expressão do mais alto nível de inteligência e significado para a transmissão de um modus de conhecimento sobre um determinado assunto (objeto informacional).

A linguagem informal, também comumente utilizada com o mesmo propósito de transmitir conhecimento ou informação, dispensa o grau de representação e significação da comunicação formal, mas eleva o nível de intimidade e cumplicidade entre os interlocutores.

Três modalidades básicas da linguagem humana movimentam ou transportam os objetos informacionais no tempo e no espaço, sendo a verbal a mais utilizada, formada por palavras, seja na escrita ou na fala. A não-verbal, basicamente representada por recursos visuais (imagens, gestos, signos corporais, desenhos), e a mista ou hibrida, representada pelo uso simultâneo das duas anteriores.

No sentido funcional, a linguagem pode ser entendida como a forma pela qual os indivíduos se manifestam em conformidade com o contexto que se inserem e para onde ou para quem está enviando a mensagem ou informação. Ao informar ou dizer algo, se espera que a maneira e a intenção da mensagem chegarão ao receptor na forma determinada pelo emissor, segundo a função utilizada. A função ou funcionalidade da linguagem alicerça-se em conjunto mínimo de elementos capaz de conduzir a informação por meio da seguinte sequência: agente emissor, quem envia a mensagem; agente receptor, a quem é dirigida a mensagem; mensagem, canal de comunicação, o meio pelo qual a mensagem é transmitida; código, o signo utilizado; e, por fim, o contexto, a condição ou circunstância na qual o objetivo ou situação em que a mensagem ou objeto informacional se insere ou a que se refere.

Esse conjunto mínimo de elementos é importante e interessa para uma melhor compreensão dos contextos seguintes, em especial no que concerne às seguintes modalidades funcionais da linguagem, assim consagradas (Everett, 2019, p. 103, 255, 273-301):

2.1.1. Função referencial

Nessa função, a linguagem é movimentada ou manejada pelo emissor de modo direto. O objetivo é transmitir com maior fidelidade possível o conteúdo do objeto informacional, em especial o de natureza científica. As principais características dessa função são a objetividade, a ênfase na informação e o foco em levar conhecimento e esclarecimento sobre o objeto limpo de juízo ou impressão do agente emissor. Caracteriza- se, principalmente, pelo sentido denotativo com uma visão isenta.

2.1.2. Função emotiva

O foco é no agente emissor ou autor que produz a mensagem. Chancela-se ao emissor a impressão ou juízo sobre o objeto, permitindo-lhe inserir seus próprios sentimentos, pensamentos e opiniões. A ideia é que o agente que movimenta ou

transmite a informação a autentique com seu próprio *modus intelligendi*. As principais características dessa função são a pessoalidade e a legitimidade para tratar a informação segundo uma visão mais intimistas sobre o objeto informacional. Na Análise do Discurso, essa função revestir-seia de um significado acima da literalidade, por chancelar autonomia em face da capacidade e da qualificação do agente emissor (Fiorin, 2009, p. 18-40).

2.1.3. Função conativa ou apelativa

Focada no agente receptor, busca pela verificação e aferição do grau de convencimento sobre o conteúdo recebido. Predomina nessa função o envolvimento do receptor com o significado dado ao objeto informacional transmitido, levando-o a adotar um determinado comportamento ou reação em face do conteúdo interpretado pelo emissor. As principais características dessa função são verbos, mormente no imperativo, que induzam a intenção emotiva, porém conotativa e apelativa, em referência à informação analisada pelo agente emissor.

2.1.4. Função metalinguística

É a função que dá ênfase à codificação da linguagem. É quando um determinado código ou signo linguístico fala por si mesmo e atribui sentido próprio ao objeto informacional. Essa função sugere uma ligação intimista do objeto informacional com termos ou vocábulos próprios e muito específicos de um determinado assunto ou área de conhecimento, a exemplo da hermenêutica do Direito. Essa função carrega em si o significado da linguagem elitista ou exclusiva de determinados ramos do conhecimento humano, por consequência a forma de manifestação de certas classes ou categorias de profissionais.

2.2. O contexto da linguística

Dois momentos históricos balizam a constituição da linguística enquanto elemento de estudo e de reflexão sobre a linguagem verbal, oral e escrita. O primeiro se passa no transcurso do século XVII com o surgimento das gramáticas gerais normativas. Naquele período, os estudos da linguagem eram fortemente marcados pelo racionalismo comunicativo. Os pensadores da época concentravam-se em analisar e estudar a língua como uma representação fiel do pensamento e buscavam demonstrar que a fala e a escrita obedeciam a princípios racionais lógicos da comunicação humana (Orlandi, 2009, p. 14-23).

O outro momento importante, já no século XIX, foi o surgimento das gramáticas comparadas. A atenção dos que trabalhavam com a linguagem volta-se para os fenômenos da transformação da linguagem no tempo, independente da vontade dos homens. As transformações seguem uma lógica adaptativa da própria língua à evolução do conhecimento humano em todos os segmentos.

A figura histórica mais expressiva relativa ao nascimento da linguística moderna foi o alemão Franz Bopp (6), considerado o precursor da linguística enquanto sistema de conjugação da língua sânscrita comparada (7). No entanto, a linguística contemporânea começa, de fato, com o suíço Ferdinand de Saussure, por meio do Curso da Linguística Geral por ele inaugurado. Mestre da Universidade de Genebra, é considerado o pai da linguística moderna.

O curso, instituído e publicado em 1916, resulta das anotações e de aulas reunidas sobre anagramas (8), hoje guardados na Biblioteca Pública de Genebra. Os anagramas formados pela transposição de letras de outras palavras ou frases, ou também formados pelo discurso poético nos versos e rimas, induzem uma linguagem sob a linguagem, capaz de produzir um efeito de anamnese no emissor ou receptor de determinada mensagem.

Orlandi (2009, p. 24-34) explica que, a partir de Saussure, a linguística concebe a língua (idioma) como objeto específico, dinâmico e determinista, à qual Saussure conceitua como um “sistema de signos”, ou seja, um conjunto de unidades independentes. Porém, conectivas entre si, capazes de formar um todo (uma ideia sobre determinado objeto informacional).

A organização interna da língua, chamada de sistema por Saussure, e, posteriormente, como estrutura, procurou valorizar a ideia de que cada elemento linguístico só adquire um valor à medida que relaciona com o todo do qual faz parte. Saussure exemplifica essa ideia comparativamente com uma peça do jogo de xadrez, por exemplo o cavalo. A identidade desta peça não depende do material de que foi feito (madeira, osso, marfim, etc.), nem mesmo de sua figura aparente (pode até ser substituído por um botão), mas o que importa, efetivamente, é a relação de oposição dele face às demais peças do xadrez, ou seja, sua identidade é determinada por sua posição no tabuleiro e seu valor no jogo.

Extraordinária a analogia desse raciocínio de Saussure, aplicada a uma compreensão da essência de um determinado objeto informacional, porque permite presumir a formação de um conjunto de ideias ou de valores intrínsecos do objeto informacional capazes de lhe atribuir sentido próprio, sem, necessariamente, ter de se utilizar pa-

lavras (chaves) ou vocabulário previamente definidos para identificação desse determinado objeto (comparativamente à figura do cavalo).

Na Espanha, a linguística documentária (9), desenvolvida para analisar de forma pormenorizada o processo linguístico relacionado ao tratamento da informação aparece no início da década de oitenta, a partir dos trabalhos de Antonio García Gutiérrez, influenciado pelos estudos de Gardin e seu contato com o Grupo Temma, da Universidade de São Paulo (USP), que teve por base a linguística estruturalista, a análise de discurso, a semiótica, a terminologia e a lógica formal para representar e organizar a informação. Segundo García Gutiérrez (1984), o objetivo fundamental da linguística, com conteúdos codificáveis e decodificáveis, divide-se em dois outros subobjetos: a estrutura da produção da informação (considera a organização e apresentação das ideias do autor); e a estrutura de representação (considera o discurso do produtor).

Miranda e Santos (2018, p. 307) observam "que no campo da linguística documentária, o trabalho é realizado em essência com informações textuais, a fim de propiciar a circulação, recuperação e uso para o público de interesse". E, que "a linguística documentária propõe critérios metodológicos para tratar a produção documentária e sua representação apoiada em referências sócias cognitivas e de linguagem das comunidades a que se destinam os produtos informacionais" (p. 307).

A linguística documentária empregada na informação de natureza específica, a terminologia, bem assim o grau de responsabilidade do profissional da informação serão verificados com mais acuidade no transcurso desta pesquisa, mas, em síntese, vale ressaltar que o trabalho do profissional ou pesquisador, enquanto leitor documentário do tema, envolve, de fato, as questões vinculadas às nuances da linguagem. Portanto, é de sua responsabilidade fazer a mediação entre essa linguagem especializada e a linguagem natural, amparado na análise de assunto documentária e suas técnicas acessórias utilizadas para organizar e representar os documentos de uma natureza muito específica.

2.3. Sujeito, sentido e ideologia

A teoria da enunciação (o eu e o outro) preconiza como centro da reflexão comunicativa o sujeito da linguagem e sua relação com o destinatário ou receptor. "Essa teoria parte da distinção entre o enunciado, já realizado, e a enunciação, que é ação de produzir o enunciado" (Orlandi, 2009, p. 58) para valorizar o processo de transferência do conteúdo informacional, ou seja, a forma e a qualidade da informação transmitida pelo sujeito emissor.

De acordo com Brandão (2004, p. 53):

> A reflexão sobre a língua tem seguido duas tendências. Segundo a epistemologia clássica, a língua tinha como função representar o real. Para ela, um enunciado era verdadeiro e correspondente a um estado de coisas existentes. Ela mobilizava, dessa forma, o conceito de verdade, privilegiando o lexicalíssimo na teorização da língua e da significação. Isto é, de acordo com essa tendência *representativa* – domínio do "dire", do nomear [...] – os nomes representariam o protótipo das categorias gramaticais, atribuindo-se ao nome o próprio ideal da representação pura. E, nesse quadro, não se colocava a questão da subjetividade.

Embora esse poder de representação da língua continue até aos dias atuais, uma nova tendência atribuiu-lhe a função demonstrativa, "o domínio do mostrar", que transcende da mera função representativa para uma nova dimensão "espessura própria", que a liberta das amarras que a prendiam a uma concepção que a enquadrava tão somente capaz de exprimir representações. Por essa nova tendência, a linguagem assume uma nova possibilidade de demonstração predicada a uma mera condição de categoria gramatical.

Nessa nova concepção teórica, o sujeito passa, então, a ocupar uma posição privilegiada e a linguagem, por seu turno, um espaço de constituição da subjetividade. Brandão (2004, p. 54) assim transcreve a incorporação dos estudos linguísticos à noção de subjetividade preconizada por Benveniste (1989, p. 82): "É o ato mesmo de produzir um enunciado, e não o texto do enunciado" – isto é, "o processo e não o produto".

Brandão sustenta que Benveniste busca o delineamento da essência da linguagem a partir das características formais da comunicação e da manifestação do sujeito. Busca, assim, por uma relação dinâmica entre a identidade e alteridade, o sujeito passa à condição de centro no espaço discursivo criado entre o emissor e o receptor (Brandão, 2004, p. 54).

Sendo o sujeito o centro do espaço discursivo, então, para Benveniste (1989, p. 83) a subjetividade passa a representar a capacidade dele (sujeito) de apresentar seu discurso por meio do exercício da língua. Esse sujeito enuncia sua posição discursiva por meio de índices formais pelos quais, em regra, a pessoalidade constitui o ponto de partida na revelação da subjetividade do assunto.

Na enunciação, o sujeito, ao instituir o "eu", requer, necessariamente, um "tu". Benveniste

(1989, p. 84) ilustra a instituição do "eu" descrevendo sobre a forma pela qual o sujeito se declara locutor da mensagem, "imediatamente, desde que ele se declara locutor e assume a língua, ele implanta o outro diante de si, qualquer que seja o grau de presença que ele atribua a este outro. Toda enunciação é explícita ou implicitamente, uma alocução, ela postula um elocutório".

Na linha de raciocínio de Brandão, "*eu e tu* são os protagonistas da enunciação e, referindo um indivíduo específico, apresenta a marca da *pessoa*. Distinguem-se, entretanto, pela marca da subjetividade: *eu* é pessoa subjetiva e *tu* pessoa não subjetiva". (Brandão, 2004, p. 56).

Brandão (2004, p. 56) afirma que, nessa correlação de subjetividade, Benveniste reconhece uma transcendência do *eu* sobre o *tu* a qual ela denomina de *ego*. Para Brandão (2004, p. 57), o ego tem sempre uma posição de superioridade em relação ao *tu*, embora nenhum dos dois se conceba sem o outro, pois são complementares e reversíveis. Seguindo o raciocínio de Brandão, o fato de o *eu* ter esse privilégio de ser único na instância discursiva faz surgir em oposição a ambos (*eu* e o *tu*) o "*ele*", que está implícito ou ausente, é não-pessoa. Embora não tendo a marca da pessoalidade, pode ao sujeito relacionar porque representa o processo que se desenvolve intrinsecamente na relação da subjetividade.

A subjetividade vai se construindo à medida em que o ego assume o centro da enunciação e se identifica como o próprio sujeito. Embora se destaque na relação discursiva a figura do parceiro "real ou imaginário", vê-se no ego o centro da enunciação. Essa atenção que ora se dedica à essa correlação "*eu/tu/ele*, onde "eu" representa o sujeito (ego); o "tu" receptor-destinatário; e o "ele" a subjetividade do enunciado (não-pessoa), em especial nesse modelo teorizado por Benveniste, na visão de Brandão (2004, p. 51-60), tem-se um particular propósito: qual seja, seguir para a narrativa sustentada na combinação das técnicas da análise documentária e no sentido holístico da análise do discurso para construção de um modelo de *indexação por sentido*.

Presumir-se-á que, tanto na indexação quanto na caracterização holística discursiva do objeto informacional, de uma forma ou de outra, o enunciado inexoravelmente vai derivar da correlação "*eu/tu/ele*". Neste contexto, o sujeito, o sentido e o contorno ideológico linguístico assumem a condição de insumo indispensável à construção desse novo modelo.

3. Discurso e construção de sentidos

Em aula ministrada no Còllege de France, em 1970, e transcrita na obra *A Ordem do Discurso*, Foucault (1996) faz um questionamento – ou melhor, uma provocação: "Mas, o que há, enfim, de tão perigoso no fato de as pessoas falarem e de seus discursos proliferarem indefinidamente? Onde, afinal, está o perigo?" (Foucault, 1996, p. 8). Antes de discorrer sobre o tema, o filósofo francês *(Ibidem,* p. 8-9) inquieta o leitor ao destacar que

> [...] em toda sociedade a produção do discurso é ao mesmo tempo controlada, selecionada, organizada e redistribuída por certo número de procedimentos que têm por função conjurar seus poderes e perigos, dominar seu acontecimento aleatório, esquivar sua pesada e temível materialidade.

Assim, se percebe uma evidente reflexão centrada nos meandros que sistematizam o processo articulador dos mecanismos de operacionalização do discurso.

O filósofo traz à luz a intrínseca relação entre o discurso e duas áreas nucleares das interações humanas: a sexualidade e a política – mais especificamente o desejo e o poder. Sendo que "o poder é o grande afrodisíaco", como certa vez afirmou o lendário e controverso político e diplomata Henry Kissinger, quando ainda era Secretário de Estado dos Estados Unidos, durante os anos de 1973–1977 (Simon, 2014). Então, entende-se que o poder não compreende o discurso como uma manifestação ou ocultação do desejo, mas como o seu próprio objeto.

Do mesmo modo, não enxerga no discurso a tradução das relações de poder, mas o próprio poder ao qual busca-se uma apropriação. Portanto, a partir dessa lógica, quem domina o discurso também domina os mecanismos de poder e os corpos, bem como atua na construção de verdades, por meio de coerções. Mais que isso: cada sociedade produz suas próprias verdades, ao delimitar, por exemplo, o que pode ou não ser dito, quem está apto a dizê-lo e ao definir sua vontade de saber, com base naquilo que a abordagem foucaultiana entende por procedimentos de interdição, segregação e vontade de verdade.

De acordo com Mello e Valentim (2021, p. 31):

> A interdição associa o discurso ao desejo de poder, o que determina que algumas palavras não podem ser ditas, ou apenas podem ser proferidas por determinados grupos que têm o direito exclusivo a certas práticas discursivas e, até ritualísticas, num determinado campo discursivo. A segregação diz respeito aos silêncios impostos pela sociedade, seja na forma de censura ou da imposição do que é moralmente aceito ou não. Por fim, a vontade de verdade que abarca os outros procedimentos, visto que em

> Foucault inexiste uma verdade em si, não há uma essência.

Nessa perspectiva, é importante considerar o discurso como um elemento semântico, visto que "[...] é uma rede de enunciados ou de relações que tornam possível haver significantes. A palavra discurso tem em si a ideia de percurso de movimento, o objeto da análise do discurso é estudar a língua em função de sentido" (Azevedo, 2013, p. 155). De tal modo, o discurso leva significado para a palavra, embutindo-lhe de textualidade, de realidade significativa, como defende Orlandi (2008). E, como produtora de sentidos para os sujeitos, "a linguagem não é, ela está" (Mira et al., 2021, p. 4).

Assim, entende-se que a linguagem é dinâmica e altamente flexível às interferências variadas de processos sociais específicos, incluindo as dimensões políticas. Santaella (2005, p. 28) reconhece a interferência histórica e cultural das tecnologias e afirma que "além de crescerem na medida exata em que cada novo veículo ou meio é inventado, as linguagens também crescem através do casamento entre meios". Com isso, fica suscetível para mudanças que agreguem incorporações de novos elementos ou sentidos estruturantes em seus repertórios linguísticos. No entanto, esse processo pode assentir viés instrumental, consciente ou inconscientemente, na formulação de enunciados para se elaborar mecanismos tácitos ou explícitos visando a construção de discurso.

Para Foucault (2008, p. 112) há coexistência enunciativa entre a intenção dos significados e as construções gramaticais que entrelaçam aspectos fundamentais do discurso, ocasionando "[...] as relações lógicas entre proposições, as relações metalinguísticas entre uma linguagem-objeto e aquela que lhe define as regras, as relações retóricas entre grupos (ou elementos) de frases". É pela estruturação do discurso que linguagem tem conotação política para determinar "[...] a questão do poder; um bem que é, por natureza, o objeto de uma luta, e de uma luta política" (Foucault, 2008, p. 136-137). Tão logo, é compreensível que a lógica discursiva nunca preza pela neutralidade ou isenção, requerendo interpretações sistemáticas do contexto. Pois, os meandros do discurso requerem a análise pormenorizada das "[...] relações sem que se tome por tema o próprio campo enunciativo, isto é, o domínio de coexistência em que se exerce a função enunciativa" (Foucault, 2008, p. 112). Afinal, o discurso tem uma finalidade comprometida com o engajamento determinado nem sempre aparente que repercute na linguagem, articulando "[...] regras de aparecimento e também suas condições de apropriação e de utilização [...]" (Foucault, 2008, p. 136-137)

Dentro dessa lógica, atravessam o tecido social, transitam por todas as instituições e, "com suas regras internas e externas, os discursos organizam e ordenam os sentidos por onde passam" (Ferreira; traversini, 2013, p. 210), a partir da posição de quem os profere. Sobre isso, Stolz (2008, p. 160) enfatiza:

> Quem diz, sempre o faz a partir de um lugar e uma intenção. Neste sentido, é importante que se tenha em mente a historicidade do discurso, a sua acomodação às diversas situações para se estabelecer, através dele, como ato impositivo, ato de verdade e de, quase sempre, ato de força.

Em outras palavras, o discurso é "um dos patamares do percurso de geração de sentido de um texto, o lugar onde se manifesta o sujeito da enunciação e onde se pode recuperar as relações entre o texto e o contexto sociohistórico que o produziu" (Gregolin, 1995, p. 17), ou seja, traz sentidos para o texto, alimenta-o, a partir do ponto de vista do sujeito da enunciação.

Mas não só isso, pois ele também está relacionado ao receptor, ao seu destino final, de modo que "todo discurso se funda sobre uma dada condição de produção que determina o modo e a forma como ele se constitui, entrando em jogo, as posições-sujeito de quem enuncia, assim como as posições-sujeito para quem o discurso está dirigido" (Coutinho, 2018, p. 239). Por isso, apresenta tamanha implicação nas relações de poder e deve ser analisado, também, diante das condições sociais e históricas sob as quais é formulado e projetado.

Foucault (2008, p. 124) oportunamente encadeia uma reflexão em que:

> Sabemos – e, talvez, desde que os homens falam – que as coisas, muitas vezes, são ditas umas pelas outras; que uma mesma frase pode ter, simultaneamente, duas significações diferentes; que um sentido manifesto, aceito sem dificuldade por todos, pode encobrir um segundo, esotérico ou profético, que uma decifração mais sutil ou apenas a erosão do tempo acabarão por descobrir; que sob uma formulação visível pode reinar uma outra que a comande, desordene, perturbe, lhe imponha uma articulação que só a ela pertence; enfim, que, de um modo ou de outro, as coisas ditas dizem bem mais que elas mesmas (Foucault, 2008, p. 124).

Identifica-se, portanto, que, para Foucault, "a questão do discurso ultrapassa o paradigma linguístico alertando que todo discurso reflete uma prática própria" (Moraes; Lima; Caprioli, 2016, p. 76), de natureza ideológica. Isso significa pensar no discurso como agente de representação cultural, em construção e regido sob as relações de

poder que circundam as diferentes esferas da sociedade.

No entanto, a abordagem foucaultiana, apesar de possuir notória relevância e presença dentro da Organização do Conhecimento, da Biblioteconomia e da Ciência da Informação, esbarra em problemáticas para a sua abordagem como método de pesquisa, não apenas associados à sua utilização, mas à sua epistemologia (Martínez-Ávila, 2012).

A principal delas é que "[...] any attempt to derive a methodology from Foucault's genealogical discourse analysis has to first deal with the author's explicit refusal to establish rules. Indeed, this is a problem that may be found in any other of his Works that organize concepts and knowledge" (Martínez-Ávila, 2012, p. 100). Além disso, falta uma abordagem epistemológica formal estabelecida pelo autor. Assim, nos nortearemos pelas contribuições pecheuxtianas de sistematização

4. A análise do discurso

A Análise do Discurso tem origem formal na escola francesa de Filosofia, por meio da prática filológica de uma conjuntura intelectual francesa que, sob a égide do estruturalismo dos anos 1960, se inscreve na articulação da linguística saussuriana, do materialismo histórico marxista e da psicanálise freudiana (Maingueneau, 2006).

A tradição de refletir e explicar os objetos textuais dos discursos inseria-se, a partir de então, no campo do saber com o encontro da prática filosófica com a prática filológica para produzir um legado instrumental metodológico de crítica textual investigativa mediante uma abordagem fundamentalmente apoiada no conjunto de vestígios, espírito, costumes e características da sociedade francesa à época marcada por uma estilística orgânica.

Para essa classe de filósofos, havia, portanto, uma necessidade de reconstruir o mundo em que surgiu o texto, relegando-se a segundo plano questões referentes às condições enunciativas mais óbvias, bem assim a linguística e o materialismo formal da filologia. Essa nova corrente de pensamento estabelece, então, um novo viés da filologia enquanto forma de estudar uma língua por meio de seus documentos escritos, que visa não só à restauração, fixação e crítica dos textos para o conhecimento do uso linguístico e sua história, mas também à compreensão de globalidade dos fenômenos culturais, especialmente os de ordem literária, a que ela serve de veículo.

Michel Pêcheux, um desses filósofos visionários da globalidade dos fenômenos culturais, em 1969, por meio da obra análise automática do discurso, no ápice do pensamento estruturalista, constituiu seus primeiros objetos discursivos, analisando-os sob a tensão da historicidade, da interdiscursividade e da sistematicidade da língua (Ferreira, 2003).

Esses objetos discursivos de Pêcheux passaram a constituir os primeiros instrumentos teóricos e metodológicos que permitiram aos analistas de assuntos incorporar as condições históricas e ideológicas em que o discurso foi produzido e, assim, experimentar gestos interpretativos e construções de sentido. Essa experiência interpretativa e a construção do sentido chancelam, então, uma espécie de permissão para ir além do conteúdo literal de um texto/discurso e possibilita uma percepção privilegiada de como ele produz e veicula sentidos, evitando reduzi-lo a algo evidente, naturalizado, hermético.

Para Pêcheux, diferentemente de Foucault, a formação discursiva possui relação direta com a formação ideológica. Enquanto na abordagem foucaultiana as relações de poder distribuem os enunciados – assim, o cruzamento é feito entre saber e poder –, na pecheuxtiana, a noção ideológica permite a compreensão e a explicação dos sujeitos particulares.

De tal forma, Paul Henry (1969, p. 12-13) observa que a ambição de Pêcheux sempre foi "abrir uma fissura teórica e científica no campo das ciências sociais, e, em particular, da psicologia social". O autor lembra que Pêcheux afirmava, por ocasião da publicação da *Análise Automática do Discurso*, que, nessa linha, estava seu fundamento profissional e principal.

Para esse fim, Pêcheux se apoiaria no que mais lhe estimulava: a problemática do Materialismo Histórico e os aspectos do grande movimento chamado Estruturalismo. Isso porque, no final dos anos 60, ocorreu o apogeu do Estruturalismo e, para Pêcheux, o que, de fato, interessava, tanto em um (o Materialismo Histórico) quanto em outro (o Estruturalismo), eram os aspectos que "supunham uma atitude não reducionista no que se refere à linguagem". Isso é ratificado pela primeira publicação de Pêcheux que dizia respeito à "situação teórica" nas ciências sociais.

Henry (1969, p. 15) concebe e desenvolve seu projeto teórico já fazendo críticas "às insuficiências do método não-linguístico da análise do conteúdo vigente nas ciências sociais à época e inaugura seu objeto teórico, o discurso, conjugando questões sobre a língua, a história e o sujeito". Segundo Pêcheux, uma teoria do discurso não pode, de forma alguma, substituir uma teoria da ideologia, nem substituir uma teoria do inconsciente, mas intervir no campo dessas teorias.

A ideia do discurso como uma produção de sentidos ao nível de um determinado contexto social, histórico e em certas condições de produção é o que caracteriza a mensagem subliminar ou indireta discursiva, conforme Orlandi (1996). A autora chancela que o funcionamento de um discurso e sua subjetividade intrínseca é: "A atividade estruturante de um discurso determinado, por um falante determinado, para um interlocutor determinado com finalidades específicas" (Orlandi, 1996, p. 197).

Segundo Orlandi (1996), a dinâmica do discurso (função do discurso) depende de dois tipos de critérios, quais sejam o de reversibilidade e polissemia. Para ela, o critério de reversibilidade se refere à interação entre os interlocutores, isto é, quanto maior esta interação e a troca de papéis entre locutor e receptor, maior a reversibilidade. O critério de polissemia, por sua vez, baseia-se na multiplicidade de significados em torno do discurso atribuídos por seus interlocutores (Orlandi, 1996, p. 29).

Partindo desses critérios, Orlandi (1996, p. 29) sugere a possibilidade de três tipos de discursos:

> No *discurso lúdico*, há a expansão da polissemia pois o referente do discurso está exposto à presença dos interlocutores; no *polêmico*, a polissemia é controlada uma vez que os interlocutores procuram direcionar, cada um por si o referente do discurso e, *finalmente* no discurso *autoritário* há a contenção da polissemia, já que o agente do discurso se pretende único e oculta o referente pelo dizer.

Nas lições de Orlandi, ancoradas nas proposições de Pêcheux, nenhum discurso se enquadra totalmente em um único tipo. O que Orlandi buscou com essa tipologia foi compreender mais a fundo como os discursos funcionam em relação às suas condições de produção e aos seus interlocutores.

Amparando-se nessas linhas teóricas de Pêcheux e Orlandi e considerando suas proposições de funcionalidade, reversibilidade, polissemia e tipos, pode-se inferir que os atos normativos jurídicos, enquanto objetos com características de discurso, podem ser classificados ou moldados como atos formais de natureza discursiva não reversível, autoritário e de polissemia contida, dada a competência exclusiva, a unipessoalidade e a impossibilidade de interação entre locutor e receptor (10).

4.1. A construção teórica da análise do discurso segundo Pêcheux

Para compreender melhor o interesse pela Análise do Discurso em muitos países influenciados pela tradição acadêmica francesa (como por exemplo Brasil ou Espanha), convém lembrar que esse tema conseguiu maior destaque nas universidades francesas nos anos 60 do século passado, período caracterizado como "primeira época", patrocinada, principalmente, pelos estudos de Michel Pêcheux. Naquele tempo, Pêcheux buscava compreender em plenitude os enunciados verbais "discursivos". A inflexão de Pêcheux pela teoria saussuriana nessa primeira fase é marcada pela relação e disputa que ele estabelece com Louis Althusser acerca do conceito de ideologia. Os objetos utilizados para a análise concentravam-se nos grandes textos políticos escritos e os dispositivos de observação se voltavam unicamente para eles.

A principal preocupação de Pêcheux e seus pares (seguidores, alunos, entusiastas) reside na questão da estruturação das mensagens (textos políticos), e essa preocupação fica bastante evidenciada na segunda parte do livro Análise automática do discurso, de Pêcheux, cujos algoritmos se voltam aos cálculos matemáticos nos quais os processos de análise automatizada do discursivo se realizava por meio da ajuda de recursos informáticos para a análise de grandes quantidades de objetos de informação (11). Ao aluno, ou outro par, cabia a responsabilidade de analisar e interpretar os dados processados, sempre pelo viés linguístico dos sentidos. Feito isto, relacionavam-se, então, os resultados, interpretações dos sentidos apoiada na Linguística, com a ideologia, com os sujeitos e com o histórico-social.

Nessa segunda época ou fase construtiva da Análise do Discurso de Pêcheux, determinados dogmas remanescentes da fase anterior passaram a ser considerados com menos rigor. Em 1975, ano da publicação de seu segundo livro, *Semântica e discurso: uma crítica à afirmação do óbvio*, Pêcheux concebeu de forma mais flexível a tese do sujeito estritamente assujeitado pela ideologia, com a formulação dos dois esquecimentos, admitindo que o sujeito possui o controle sobre os enunciados verbais que emite – "essa relação sujeito e enunciado" guarda perfeita correlação com a questão da linguagem própria e da hermenêutica do Direito em relação aos enunciados (atos, normas, objetos informacionais jurídicos) produzidos pelos respectivos profissionais do Direito.

Na visão de Pêcheux (2014, p. 161-163), os atos e comportamentos do sujeito, tais como consciência e atividade, são as fontes que determinam "sua realidade" e tendem a seguir uma repetição à qual ele denominou de "mito idealista da interioridade". Nesse "mito", aquilo que foi dito não poderia ser diferente do já dito, pois é aí que o sujeito deve encontrar uma reflexão sobre sua "verdade" e sobre si mesmo. Para Pêcheux, é nesse

mito idealista que se assenta a formação discursiva. Para tanto, Pêcheux baseia-se na oposição de Freud sobre o "sistema pré-consciente" e o "sistema consciente" para estabelecer dois tipos radicalmente diferentes de "esquecimentos" inerentes ao discurso.

A teoria do Primeiro Esquecimento:

> [...] a noção de "sistema inconsciente" para caracterizar um outro "esquecimento", o esquecimento nº 1, que dá conta do fato que o sujeito falante não pode, por definição, se encontrar no exterior da formação discursiva que o domina. Nesse sentido, o esquecimento nº 1 remetia, por uma analogia com o recalque inconsciente, a esse exterior, na medida em que – como vimos – esse exterior determina a formação discursiva em questão.

A teoria do Segundo Esquecimento:

> [...] "esquecimento" pelo qual todo sujeito-falante "seleciona" no interior da formação discursiva que o domina, isto é, no sistema de enunciados, formas e sequências que nela se encontram em relação de paráfrase – um enunciado, forma ou sequência e não um outro, que, no entanto, está no campo daquilo que poderia reformulá-lo na formação discursiva considerada.

Em seu livro *O discurso: estrutura ou acontecimento*, Pêcheux foca em um enunciado político comum: "On a gagné", adotado e repetido pelos eleitores de François Mitterrand, do partido de esquerda, vencedor das eleições para presidente da República Francesa, em 1981. Esse enunciado, segundo Pêcheux (2015, p. 24), "é atravessado por discursividades da mesma maneira que os escritos doutrinários, pois revela uma estrutura 'On a gagné' como sujeito indefinido referindo-se indeterminadamente aos militantes do partido esquerdista francês ou ao povo geral da França".

Pêcheux (2015) afirmava que a ausência ou aparente sentido vazio do discurso-texto é um requisito constitutivo da linguagem que aparece sob a forma de variados elementos: negação, hipótese, desejo, subjuntivo, formas de presente/passado/futuro, imperativo, "eu" diferenciando-se de "nós", a alteridade encontrada em "ele(s)" e "ela(s)". Ao endosso dessa afirmativa, Pêcheux atribuía as seguintes analogias ou vocábulos abstrativos correspondentes como: "o povo", "as massas", "o proletariado", "a luta de classes".

Essas analogias podiam ser mostradas (pintadas, filmadas ou televisionadas) enquanto conceitos, porém, como disfarces subliminares – na análise análoga –, Pêcheux questiona se, de fato, uma abstração pode ser pintada sem disfarces. Sabia ele que as referências "o povo" e "as massas" representavam os conceitos supremos da teoria marxista. Logo, as questões levantadas por Pêcheux, por meio das mensagens cifradas ou "codificadas", revelava ou induzia para uma compreensão da objetividade dos preceitos marxistas manipulados, principalmente, pelos políticos e, por consequência, remete a uma revisão das bases do projeto teórico da análise do discurso, pois as abstrações, sofriam influência de disfarces, do inconsciente, do simbólico.

Como últimas manifestações de um autor inquieto, Michel Pêcheux, que suicidou-se em 1983, em seus últimos textos, demonstrava aflição em relação às transformações do discurso político, embora não se detivesse ou se preocupasse com as manipulações das tecnologias de comunicação de massa e futuras consequências de sua popularização aos quais já eram perceptíveis no início da década de 80, como instrumentos de manobras dos homens públicos.

4.2. Atuais orientações conceituais da teoria do discurso

O discurso, enquanto "língua objeto de mensagem", segundo Saussure, é:

> a) [...] a parte social, a linguagem, exterior ao indivíduo, que por si só não pode nem criá-la nem modificá-la, e:
>
> b) [...] a língua é uma instituição social; mas se distingue, por vários traços, das outras instituições políticas, jurídicas etc. Para compreender sua natureza especial, uma nova ordem de fatos precisa intervir. A língua é um sistema de signos que exprime ideias, e por isto comparável à escrita, ao alfabeto dos surdos mudos, aos ritos simbólicos, às formas de polidez, aos sinais militares etc. Ela é somente o mais importante desses sistemas. Pode-se, pois, conceber uma *ciência que estuda a vida dos signos no seio da vida social*, ela formaria uma parte da psicologia social e consequentemente da psicologia geral [...] (Saussure apud Pêcheux, 1969, p. 69, itálicos do autor).

Essa segunda orientação conceitual – relativa ao objeto e sua dependência com outros objetos situados no mesmo plano – opera, segundo Saussure, uma dupla divisão: alia-se ao sistema semiológico (a língua), que é pensado como um estatuto científico potencialmente equivalente, e entra no campo da teoria regional do significado. A outra oposição que é evocada por Saussure, por meio do termo *instituição*, é: ela permite separar os sistemas institucionais jurídico, político etc. da série dos sistemas institucionais semiológicos, e excluí-los simplesmente do campo da teoria do significante regional.

Assim, a língua (discurso) é pensada por Saussure (Pêcheux, 1969, p. 69) como um objeto científico homogêneo (pertencente à semiologia), cuja especificidade se estabelece sobre duas principais exclusões teóricas:

- A exclusão da *fala* no inacessível da ciência linguística;

- A exclusão das instituições "*não semiológicas*" para fora da zona de pertinência da ciência linguística.

Essa perspectiva de Saussure em referência à língua-discurso se estendeu para os mais diversos seguimentos do conhecimento humano, incluindo numerosos e relevantes estudos na área da Informação e Comunicação que se baseiam na análise do discurso de matriz francesa (Castanha et al. 2017). O artigo *Política de indexação e seus sentidos: um estudo a partir da Análise do Discurso*, dos autores Garcia, Redigolo, Barros e Moraes (2019) é de especial importância para a verificação da especificidade de domínio deste estudo porque os autores promovem uma espécie de *background* sobre a temática da política de indexação e a maneira como ela está sendo discutida na atualidade. Nesse estudo, Garcia, Redigolo, Barros e Moraes analisam quatro artigos de diferentes autores, recortados temporalmente nos anos de 2011-2016 (Fujita et al., 2012; Fujita e Santos, 2016; Lousada et al., 2011; Silva e Boccato, 2012) e apresentam suas percepções em relação ao aperfeiçoamento dos processos e procedimentos para a atual representação e recuperação da informação. Também de especial relevância porque analisam, nos quatro artigos, os aspectos relacionados à interação da indexação com a Análise do Discurso de Pêcheux, para a melhoria do processo de indexação. Essa conexão de assuntos justifica o breve resumo do artigo.

Garcia, Redigolo, Barros e Moraes observam que não há pretensão de efetuar uma análise exaustiva sobre os sentidos e aspectos envolvidos dos vários discursos sobre política de indexação, mas indagam como se caracteriza a atual política de indexação, o que estão dizendo os autores e, principalmente, "qual o sentido expresso nos enunciados proferidos por seus sujeitos" (p. 170). Lembram que, para Barros (2017), tanto o discurso oral, quanto o escrito podem transmitir diferentes significados (12).

Para Garcia, Redigolo, Barros e Mores (2019), à medida em que os modelos sociais evoluem e passam por mudanças, os procedimentos vinculados à produção e à disseminação da informação também se transformam e tornam-se mais dinâmicos. Daí que o aparecimento de novos métodos e técnicas chegam com o propósito de aperfeiçoar a organização do conhecimento para fins de armazenagem e recuperação. A indexação, enquanto técnica, constitui ferramenta fundamental, em especial para as bibliotecas que priorizam os conceitos-assuntos dos objetos informacionais para manejar as coleções, por meio da valorização temática para disponibilizá-los em catálogos para fins de recuperação para os usuários.

Garcia, Redigolo, Barros e Moraes (2019) salientam que política de indexação é uma conduta (ferramenta) que deve ser utilizada, sobretudo, para conduzir um eficaz tratamento e representação temática da informação por meio da indexação. Todavia, para eles, existem poucos estudos direcionados a esse segmento, e citam as palavras de Rubi (2004): "Poucos autores trabalham com a política de indexação" (Rubi, 2004, p. 12 apud Garcia; Redigolo; Barros e Moraes, 2019, p. 173). É neste contexto de escassez que Garcia, Redigolo, Barros e Moraes (2019) intentam verificar "como a política de indexação vem sendo caracterizada na literatura de artigos científicos atuais através dos discursos dos sujeitos que tratam sobre tal temática" (p. 173), bem assim, "quais os sentidos expressos nesses discursos a partir dos contextos, condições de produção e ideologias" (p. 173). Esclarecem que, para isto, farão uso da teoria da Análise do Discurso focada especialmente nas considerações de Barros (2015), quando ele afirma que essa teoria se preocupa com o além-texto, quando verifica "em que medida a construção de um texto remete às esferas ideológicas" (Barros, 2015, p. 69 apud Garcia; Redigolo; Barros e Moraes, 2019, p. 173) e, ainda, na concepção de Orlandi, que entende que:

> O texto é a unidade de análise afetada pelas condições de produção e é também o lugar da relação com a representação da linguagem: som, letra, espaço, dimensão, direcionada, tamanho. Mas é também, e sobretudo, espaço significante: lugar de jogo de sentidos, de trabalho da linguagem, de funcionalidade da discursividade. Como todo objeto simbólico, ele é objeto de interpretação. O analista tem de compreender como ele produz sentidos, o que implica em saber, tanto como ele pode ser lido, quanto como os sentidos estão nele. Na análise de discurso, não se toma o texto como ponto de partida absoluto (dadas as relações de sentido) nem de chegada. Um texto é só uma peça de linguagem de um processo discursivo bem mais abrangente e é assim que deve ser considerado. Ele é um exemplar do discurso. [...] Não é sobre o texto que falará o analista, mas sobre o discurso. (Orlandi, 2020, p. 70).

Na sequência, Garcia, Redigolo, Barros e Moraes (2019) procedem à metodologia e à análise textual dos quatro artigos selecionados e concluem, em especial com o objeto desta pesquisa, que:

> Ao buscar no discurso os prováveis sentidos que a política de indexação pode assumir atualmente,

> considerando o sujeito, a sua história, a ideologia e o seu contexto social, verificou-se, primeiramente, que os discursos atuais sobre políticas de indexação são constituídos pelas formações discursivas dos ambientes onde seus sujeitos estão inseridos, ambientes acadêmicos das instituições de ensino, mais precisamente nos programas de pós-graduação e seus grupos de pesquisas, espaços historicamente associados à produção de conhecimento [...] (Garcia; Redigolo; Barros e Moraes, 2019, p. 184).

Para além disso, os autores ainda destacam:

> [...] os discursos produzidos nesses ambientes sobre políticas de indexação como guia e também como filosofia refletem as condições de produção e seus contextos institucionais, nos quais os sujeitos buscam por estudos mais aprofundados para a temática visando à melhoria dos processos de indexação em um ideal de qualificação constante, tendo a política não só como manual de orientação de processos, mas também como algo que reflete a natureza e cultura organizacional da unidade de informação quanto à organização e à disseminação da informação (Garcia; Redigolo; Barros e Moraes, p. 185).

Na área da informação jurídica, por exemplo, Reis (2019), apresenta uma especial similaridade com a proposição da *indexação por sentido*, na medida em que aborda os aspectos da semiose e da inferência da estrutura textual da doutrina jurídica, pois, de qualquer forma, verifica sobre as questões das singularidades da linguagem do Direito, a exemplo da hermenêutica e das expressões dogmáticas no contexto do espaço tempo de geração do objeto informacional em análise. Reis demonstra essa preocupação ao dizer que:

> Cada profissional que efetua a prática da leitura documentária é único, como consequência disso, a análise do documento nunca ocorrerá da mesma forma. Vários fatores devem ser levados em conta, quando se estuda o processo de leitura documentária feito por profissionais da informação, como estratégia de leitura, conhecimento prévio, domínio de atuação e tipo de estrutura do documento analisado (Reis, 2019, p. 9).

Reis (2019) confirma essa narrativa afirmando que há necessidade de avanço nos estudos dos processos metacognitivos na leitura documentária realizada pelos bibliotecários jurídicos – afirma que é preciso de apoio nas teorias associadas à construção de significados, de forma a verificar os aspectos relacionados à semiótica, à abdução, dedução e indução. Este exemplo poderia ser aplicado a muitos outros campos relacionados com a Informação e a Comunicação onde a indexação por sentido e a analise do discurso possam contribuir ao trabalho no domínio.

No campo das bibliotecas e a organização do conhecimento, por exemplo, o trabalho de Larissa de Mello Lima (2021) teve como foco encontrar uma forma fidedigna de representar as singularidades do texto narrativo de ficção do gênero conto, no contexto da indexação de obras de ficção. O cenário encontrado, que consistiu no problema de pesquisa, foi a percepção do tratamento superficial dado aos textos literários no momento da análise documental, culminando em representações que se centravam nos aspectos externos e formais do documento, enquanto o assunto era confundido com as categorias gênero e nacionalidade, por exemplo. Acredita-se que este fenômeno se dava pela incompatibilidade de aplicação das diretrizes do texto científico ao texto narrativo de ficção do gênero conto. O objetivo principal do estudo foi criar um modelo de leitura direcionado para o conteúdo do texto literário, levando em conta as peculiaridades do texto narrativo de ficção do gênero conto com base na autora Clarice Lispector. Para a criação do modelo de leitura, foi necessário se apoiar em uma metodologia robusta: a Análise do Discurso de matriz francesa oferecendo os aportes teóricos e a Análise do Discurso literário, os conceitos. Por meio destas perspectivas que foram elucidadas ao longo deste trabalho, foi possível entender que o discurso literário de Clarice Lispector existe diante da opacidade do dizer, subvertendo-o, pois, os conceitos como formação discursiva e interdição de Foucault e Pêcheux nos explicam que “não é possível dizer tudo em qualquer circunstância”. Perspectiva esta que Clarice Lispector rompe no conto “Ruído de passos”, ao falar sobre masturbação feminina, valendo-se de eufemismos, por exemplo. O modelo de leitura não foi criado com a pretensão de ter um caráter prescritivo porque não reflete a perspectiva crítica da Análise do Discurso e da Análise do Discurso Literário que deram base teórica e metodológica para a criação do modelo. Acredita-se que este trabalho esboce uma nova linha narrativa crítica para os estudos da organização do conhecimento e para os bibliotecários/indexadores, no momento de realizar a leitura documental e que o modelo de leitura tem o potencial de ser uma maneira viável de representar os textos narrativos de ficção do gênero conto.

No artigo *Análise do discurso e ciência da Informação: aportes teóricos para organização e representação da Informação*, de João Batista Ernesto Moraes, Larissa Mello Lima e Mariana Silva Caprioli (2016), uma perspectiva conceitual e teórica da escola francesa de Análise do Discurso é apresentada como uma metodologia complementar e válida para ser utilizada em estudos teóricos de organização e representação da informação em Ciência da Informação. São selecionados autores como Foucault (2010,

1986), Orlandi (1999) e Mazière (2007), que oferecem debates fundamentais sobre a Análise do Discurso. O segundo passo é discutir a questão da interdisciplinaridade, com foco nas Ciências da Informação, a fim de fortalecer as bases para o estabelecimento da relação entre a Análise do Discurso da escola francesa e a CI.

Lima, Moreira e Moraes (2016), no artigo *Linguística documentária e Análise do Discurso: um mapeamento entre conceitos*, apresentam um delineamento de conceitos equivalentes entre a Linguística documentária e a Análise do Discurso. Para tanto, utilizam um estudo teórico exploratório como metodologia, a fim de construir mapas conceituais para identificar conceitos chave tanto da Análise do Discurso quanto da linguística documentária, tendo, assim, fins comparativos. Lima, Moreira e Moraes partem da seguinte problemática: existem relações teóricas conceituais entre a Análise do Discurso e a Linguística Documentária? Para solucionar tal questão, o objetivo geral deste trabalho é fornecer um panorama verticalizado acerca da possível relação conceitual entre ambas. No que tange aos objetivos específicos, o trabalho buscou, primeiramente, identificar definições conceituais de cada campo, para que se torne possível, em um segundo momento, selecionar os conceitos que irão formar os mapas conceituais de ambos. Assim, em um terceiro momento, é possível visualizar os conceitos comuns ou aparentemente comuns, os conceitos complementares e os que se contrapõem. Como resultado, é possível sinalizar que existe consonância entre os conceitos de ambas as áreas, perspectiva que pode abrir portas para novos estudos mais aprofundados que auxiliem na organização e representação do conhecimento.

Morris (2010) aborda, no estudo *Individual Differences in the Interpretation of Text: Implications for Information Science*, a relevância das questões das diferenças individuais (sujeitos) nas produções textuais e destaca que a maioria das tarefas realizadas ao nível da Ciência da Informação, em especial na Biblioteconomia, a exemplo da indexação e classificação, exigem atenção do leitor profissional na interpretação dos significados textuais para as nuances e aspectos relacionados às diferenças individuais de quem os produz. Considera que é possível perceber e modelar diferenças por meio de análise do perfil da semântica lexical, a partir das características individuais.

Morris desenvolve o estudo baseando-se no pressuposto de que as palavras são elementos de natureza pessoal, com significância e sentidos vinculados ao autor, e que podem ser analisadas e interpretadas na verificação da coesão lexical, a qual apresenta características distintas em razão do sujeito e do contexto. Nos termos conclusivos do estudo, um detalhe de interessante pertinência com o tema desta pesquisa é a dificuldade para automação de variáveis abstratas a exemplo das características individuais semânticos-lexical; do mesmo modo que seria a automação de uma *indexação por sentido*. Entretanto, Morris, apresenta uma provável solução em relação à sua perspectiva teórica:

> The major implication of recognizing individual differences in text interpretation is to model them computationally. Rather than viewing them as a problem to be overcome, they can be viewed as a natural aspect of interpretation. Thus, by studying the details of individual differences of various aspects of text meaning, such as lexicals cohesion, we can attempt to create models of them. This way, the computer could interpret a text differently for each reader. A very useful application of such models would be for information retrieval. (2010, p. 147).

Morris (2010) aconselha que, ao estudarem as características e os detalhes das diferenças ou marcas individuais sobre os vários aspectos da produção autoral de um dado texto, como a coesão lexical, por exemplo, poder-se-iam criar modelos padrão a partir desses aspectos. Assim, seria possível o desenvolvimento de um software para processar, reconhecer e vincular os objetos informacionais aos seus respectivos autores. Para Morris, essa seria uma solução eficiente e ideal para recuperação de informação, baseada nos reconhecimentos das diferenças individuais e na modelagem dos textos em relação aos seus autores. Factível ou não, essa perspectiva interessa de perto para a temática deste estudo, pois, de alguma forma, o ponto de vista de Morris está relacionado com a ideia de “sentido”. Embora não seja um sentido na forma preconizada por Pêcheux, tem o mérito de se buscar compreender a mensagem, ainda que pelas características individuais de quem produz o discurso (texto).

A priori, estes são os estudos no Brasil que mais se aproximam e demonstram interesse mais direto sobre a recuperação de objeto informacional de natureza específica, ou que, pelo menos, de alguma forma, teorizam sobre as possiblidades e necessidade de inovação e renovação de modelos e técnicas da Ciência da Informação (e áreas afins como a Biblioteconomia), em especial para a indexação. Fora do Brasil, os trabalhos que discutem ou aplicam a analise de discurso de matriz francesa tem focado mais no trabalho de Foucault (Budd e Raber, 1996;

Fora do Brasil, a análise do discurso, na perspectiva de Foucault, tem sido trabalhada extensivamente na Ciência da Informação, desde os anos

90 (ex. Frohmann, 1992; 1993; 1994; 2001; Budd e Raber, 1996; Radford, 2003; Radford e Radford, 2005; Budd, 2006; Andersen e Skouvig, 2006; Haider e Bawden, 2007; Olsson, 2010; Martínez-Ávila, 2012; Moulaison et al. 2014; Martínez-Ávila e Fox, 2015; Martínez-Ávila et al., 2015). Enquanto no Brasil Michel Foucault apresenta junto a Michel Pêcheux a maior frequência de cocitação nos trabalhos sobre análise do discurso na Ciência da Informação (Castanha et al., 2016), fora do Brasil (ou da França) Pêcheux é pouco conhecido e citado na Ciência da Informação, e, quando tem sido citado (ex. Haider e Bawden, 2007), é feito de forma anedótica.

5. Conclusão

Para finalizar a proposta e breve panorâmica sobre a especificidade de domínio deste estudo, restauram-se aqui as disposições da primeira premissa, na qual foi estabelecido que uma leitura documentária eficaz sobre um determinado tema muito específico é um grande desafio da atualidade para o profissional da informação. Considerando que esse desafio pode ser mitigado com o desenvolvimento de mais e melhores estudos que buscam a compreensão e a desmistificação do problema da recuperação de informação de natureza muito específica. Então, deve-se ter em conta um retrato das proposituras, tendências, padrões mais recentes; bem assim, o que estão pensando os agentes, a partir de quais contextos são produzidos os sentidos de seus discursos e como estão se relacionando no campo deste tema na especificidade de domínio.

Notas

(1) *Modistae*: movimento reconhecido como uma escola filosófica de gramática denominada de gramática modista ou especulativa presente na França, Alemanha, Inglaterra e Dinamarca nos séculos XIII e XIV. Vitalizou se em oposição à gramática pedagógica (Bursill-Hall, 1972).

(2) *Modus*: aquilo que existe em si mesmo também pode existir na mente. Um modo não é "um que", mas "um como". Algo que pode existir em diferentes símbolos, porém sem mudar a compreensão. Por exemplo o que existe em si mesmo pode existir também na mente (a figura de um cavalo pode ser escrita ou representada em diferentes línguas, porém com um mesmo significado na mente) (Bursill-Hall, 1972).

(3) *Essendi*: é o modo pelo qual o alguam "o que" existe em si mesmo. Independe de qualquer ação ou vontade exógena. Está inserido na própria existência primária, é o sujeito. É um modo independente, embora a entidade possa exigir ou depender de muitos fatores ambientais e circunstâncias para a manutenção de sua existência sua imagem e subjetividade permanecerá oriundos de seus princípios internos e sua essência (Bursill-Hall, 1972).

(4) *Intelligendi*: é a maneira pela o que existe por si mesmo (seja em si mesmo, ou em outro) existe na mente. É aquilo considerado a respeito de um ou de outro por semelhança ou por compreensão. É o que existe "in" no juízo do interlocutor ou do intérprete, a partir do modus (Bursill-Hall, 1972).

(5) *Significandi*: assim como o que existe em si mesmo, em outro ou entre dois outros, também pode existir na mente inteligente na forma de signo (símbolo ou sinal). Um modo específico de entendimento de um em relação ao(s) outro (s). Pode explicar e transmitir o que existe no entendimento, visto que depende de um (Bursill-Hall, 1972).

(6) Franz Bopp, nascido em Mainz, foi um Filólogo linguista alemão e professor de filologia e sânscrito na Universidade de Berlim. Demonstrou a importância do sânscrito para as línguas indo-européias e é considerado o fundador da linguística comparativa. Seu talento apareceu inicialmente em Über das Conjugationssystem der Sanskritsprache (1816). (https://educacao.uol.com.br/biografias/franz-bopp.htm).

(7) Sânscrito - diz-se de ou grupo de línguas indo-arábicas, antigas e modernas, que formam a maioria das línguas indo-europeias da Índia, Paquistão, Bangladesh e outros países vizinhos. O sânscrito ou língua sânscrita é uma língua ancestral do Nepal e da Índia. Embora seja uma língua morta, o sânscrito faz parte do conjunto das 23 línguas oficiais da Índia, porque tem importante uso litúrgico no hinduísmo, budismo e jainismo. (https://educacao.uol.com.br/biografias/franz-bopp.htm).

(8) Os exemplos mais comuns dos anagramas surgidos da teoria de Saussure são a subposição das letras de Iracema e a palavra América em José de Alencar; e o verso latino "Mors perfecti tua ut essent" que evoca as vogais do nome Cornelius sem que ele seja explicitado.

(9) No Brasil, o termo consolidado é linguística documentária, devido à sua ligação com a análise documentária e as influências francesas experimentadas por pesquisadores brasileiros. O termo linguística documental foi formulado por García Gutiérrez, nos anos 80, na Espanha (García Gutiérrez, 1984). Para este estudo, será adotado o termo linguística documentária.

(10) A referência aos critérios da reversibilidade e da polissemia (fundamento teórico de Pêcheux e Orlandi) como elementos do discurso aplicada à informação jurídica será trabalhada com melhor foco visando-se inferir os efeitos desses fenômenos na hermenêutica do Direito e na recuperação da informação.

(11) Nesse período, Pêcheux e sua equipe, além dos textos políticos, analisavam também os discursos.

(12) Garcia, Redigolo, Barros e Moraes (2019) esclarecem que no entendimento de Barros (2017) alguns discursos "podem ser claramente percebidos por seus receptores enquanto outros podem estar implícitos nas manifestações textuais ou orais." (Barros, 2017 apud Garcia, Redigolo, Barros & Moraes, 2019, p.170).

Referencias

Andersen, Jack; Skouvig, Laura (2006). Knowledge organization: a sociohistorical analysis and critique. // Library Quarterly. 76:3 (jul. 2006) 300-22. https://doi.org/10.1086/511139

Azevedo, Sara Dionizia Rodrigues de (2013). Formação discursiva e discurso em Michel Foucault. // Filogênese. 6:2 (jul./dez 2013) https://www.marilia.unesp.br/Home/RevistasEletronicas/FILOGENESE/saraazevedo.pdf.

Barros, Thiago Bragato (2017). Discurso, informação e conhecimento: perspectivas iniciais à Ciência da Informação. // Brazilian Journal of Information Studies: Research

Trends. 11:3 (out 2017) 26-33. https://doi.org/10.36311/1981-1640.2017.v11n3.04.p26

Benveniste, Émile (1989). Problemas de linguística geral II. Tradução de Eduardo Guimarães et. al. Revisão técnica da tradução Eduardo Guimarães. Campinas: Pontes, 1989.

Brandão, Helena Hathsue Nagamine. (2004). Introdução a análise do discurso. (2. ed.) Campinas: Editora da Unicamp, 2004.

Budd, John (2006). Discourse analysis and the study of communication in LIS. // Library Trends. 55:1 (summer 2006) 65-82. https://doi.org/10.1353/lib.2006.0046

Budd, John M.; Raber, Douglas (1996). Discourse analysis: method and application in the study of information. // Information Processing & Management. 32:1 (march 1996) 217-226. https://doi.org/10.1016/S0306-4573(96)85007-2

Castanha, Renata Cristina Gutierres; Lima, Larissa de Mello; Martínez-Ávila, Daniel (2017). Análise do discurso sob a perspectiva bibliométrica nos estudos de ciência da informação no Brasil. // Perspectivas em Ciência da Informação. 22:1 (2017) 17-37, 2017. http://dx.doi.org/10.1590/1981-5344/2813

Coutinho, Renata Corrêa (2018). Publicidade e discurso: um gesto de leitura sobre o discurso publicitário. // Revista Eletrônica Internacional de Economia Política da Informação, da Comunicação e da Cultura. 20:2 (maio/ago. 2018) 236-245.

Everett, Daniel L. (2019). Linguagem: a história da maior invenção da humanidade. São Paulo: Contexto.

Fernandes, Cleudemar Alves (2005). Análise do Discurso: reflexões introdutórias. Goiânia: Trilhas Urbanas, 2005.

Ferreira, Maria Cristina Leandro (2003). O caráter singular da língua na análise do discurso. // Organon: Revista do Instituto de Letras da UFRGS. 17:35 (2003) 190-200. https://doi.org/10.22456/2238-8915.30023

Ferreira, Mauricio dos Santos; Traversini, Clarice Salete (2013). A Análise Foucaultiana do Discurso como Ferramenta Metodológica de Pesquisa. // Educação & Realidade. 38:1 (jan/mar 2013) 207-226.

Fiorin, Jose Luiz. (2009). Elementos de análise do discurso. (14. ed.). São Paulo: Contexto.

Foucault, Michel (1986). A arqueologia do saber. Rio de Janeiro: Forense, 1986.

Foucault, Michel (1996). A ordem do discurso. 3. ed. São Paulo: Loyola, 1996.

Foucault, Michel (2010). A ordem do discurso. São Paulo: Loyola, 2010.

Frohmann, Bernd (1992). The power of images: a discourse analysis of the cognitive viewpoint. // Journal of Documentation. 48:4 (april 1992) 365-386. https://doi.org/10.1108/eb026904

Frohmann, Bernd (1994a). Discourse analysis as a research method in library and information science. // Library and Information Science Research. 16:2 (spring 1994) 119-138. https://doi.org/10.1016/0740-8188(94)90004-3

Frohmann, Bernd (1994b). Communication technologies and the politics of postmodern information science. // Canadian Journal of Information and Library Science. 19:2 (1994) 1-22.

Frohmann, Bernd (2001). Discourse and documentation: some implications for pedagogy and research. // Journal of Education for Library & Information Science. 42:1 (winter 2001) 13-26. https://doi.org/10.2307/40324034

Fujita, Mariângela Spotti Lopes; Agustín Lacruz, María del Carmen; Gomez Diáz, Raquel (2012). A situação atual da indexação nas tarefas bibliotecárias. // Perspectivas em Ciência da Informação. 17:1 (2012) 94-109. https://doi.org/10.1590/S1413-99362012000100006

Fujita, Mariângela Spotti Lopes; Santos, Luciana Beatriz Piovezan dos (2016). Política de indexação em bibliotecas universitárias: estudo diagnóstico e analítico com pesquisa participante. // TransInformação. 28:1 (2016) 59-76. http://dx.doi.org/10.1590/2318-08892016002800005

Garcia, Valdenise César; Redigolo, Franciele Marques; Barros, Thiago Henrique Bragato; Moraes, João Batista Ernesto de (2019). Política de indexação e seus sentidos: um estudo a partir da Análise do Discurso. // Informação & Informação. 24:1 (2019) 169-189. https://doi.org/10.5433/1981-8920.2019v24n1p169

García Gutiérrez, Antonio Luis (1984). Linguística documental: aplicación a la documentación de la comunicación social. Barcelona: Editorial Mitre, 1984.

Gregolin, Maria do Rosario Valencise (1995). A análise do discurso: conceitos e aplicações. // ALFA: Revista de Linguística. 39 (1995). https://periodicos.fclar.unesp.br/alfa/article/view/3967.

Haider, Jutta; Bawden, David (2007). Conceptions of 'information poverty' in LIS: a discourse analysis. // Journal of Documentation. 63:4 (2007) 534-557. https://doi.org/10.1108/00220410710759002

Henry, Paul (1969). Os fundamentos teóricos da "análise automática do discurso" de Michel Pêcheux. // Gadet, F.; Hak, T. (eds.). Por uma análise automática do discurso: uma introdução à obra de Michel Pêcheux. Campinas: Editora da Unicamp, 2014. 11-38.

Lima, Larissa de Mello (2021). Modelo de Análise documental de textos literários pela perspectiva da análise do discurso: um estudo dos contos de Clarice Lispector. [tese doutorado]. Marília: Universidade Estadual Paulista, 2021. http://hdl.handle.net/11449/204441

Lima, Larissa Mello; Moreira, Walter; Moraes, João Batista Ernesto (2016). Linguística documentária e Análise do Discurso: um mapeamento entre conceitos. // Seminário em Ciência da Informação: fenômenos emergentes em Ciência da Informação, Londrina-PR, Brasil. http://www.uel.br/eventos/cinf/index.php/secin2016/secin2016/paper/viewFile/328/183

Lousada, Mariana; Lopes, Elaine Cristina; Fujita, Mariângela Spotti Lopes; Valentim, Marta Lígia Pomim (2011). Políticas de indexação no âmbito da gestão do conhecimento organizacional. // Informação & Sociedade: Estudos. 21:1 (2011) 191-202

Maingueneau, D. (2006). Discurso literário. Tradução de Adail Sobral. São Paulo: Contexto, 2006.

Martínez-Ávila, Daniel (2012). Problems and Characteristics of Foucauldian Discourse Analysis as a Research Method. // Smiraglia, Richard P.; Lee, Hur-Li (eds.). Cultural Frames of Knowledge. Würzburg: Ergon-Verlag, 2012. 99-110.

Martínez-Ávila, Daniel; Melodie J. Fox (2015). The Construction of Ontology: A Discourse Analysis. // Smiraglia, Richard P.; Lee, Hur-Li (eds.). Ontology for Knowledge Organization. Würzburg, Germany: Ergon, 13-37.

Martínez-Ávila, Daniel; Smiraglia, Richard; Lee, Hur-Li; Fox, Melodie (2015). What Is an Author Now? Discourse Analysis Applied to the Idea of an Author. // Journal of Documentation. 71:5 (2015) 1094-1114. https://doi.org/10.1108/JD-05-2014-0068

Mazière, Francine. (2007). A análise do discurso: história e práticas. São Paulo: Parábola Editora, 2007.

Mazolla, Renan Belmonte (2009). Análise do Discurso: um campo de reformulações. // Milanez, Nilton; Santos, Janaina de Jesus (eds.). Análise do Discurso: sujeito, lugares e olhares. São Carlos: Claraluz, 2009. 7-16.

Mello, Mariana Rodrigues Gomes de; Valentim, Marta Lígia Pomim (2021). Análise do discurso: diálogos epistemoló-

gicos em Foucault e Heidegger. // Logeion: filosofia da informação. 7:2 (2021). 24-43. https://doi.org/10.21728/logeion.2021v7n2.p24-43

Mira, Bianca Savegnago de; Farias, Mary Elizabeth Sampaio de Oliveira; Brito, Jean Fernandes; Guaraldo, Tamara de Souza Brandão (2021). Nas trilhas dos sujeitos discursivos. // Informação em Pauta. 6:00 (2021) 1-16. https://doi.org/10.36517/2525-3468.ip.v6i00.2021.60238.1-16

Miranda, Roseli; Santos, Cibele Araújo Camargo Marques dos (2018). Documentação jurídica: interfaces da leitura documentária, linguagem e análise do discurso no tratamento da informação. // RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação. 16:3 (2018) 299-316. https://doi.org/10.20396/rdbci.v16i3.8650313

Moraes, João Batista Ernesto; Lima, Larissa Mello; Caprioli, Mariana Silva (2016). Análise do discursoe ciência da informação: aportes teóricos para organização e representação da Informação. // Scire. 22:2 (jul.-dic. 2016) 75-85.

Morris, Jane (2010). Individual differences in the interpretation of text: implications for information. // Journal of the American Society for Information Science and Technology. 61:1 (2010) 141-149. https://doi.org/10.1002/asi.21222

Moulaison, Heather Lea; Dykas, Felicity; Budd, John M. (2014). Foucault, the Author, and Intellectual Debt: Capturing the Author-Function Through Attributes, Relationships, and Events in Knowledge Organization Systems. // Knowledge Organization. 41:1 (2014) 30-43. https://doi.org/10.5771/0943-7444-2014-1-30

Olsson, Michael R. (2010). Michel Foucault: Discourse, Power/Knowledge, and the Battle for Truth. // Leckie, Gloria J.; Given, Lisa M.; Buschman, John E. (eds.). Critical theory for library and information science: exploring the social from across the disciplines. Santa Barbara: Libraries Unlimited, 63-74.

Orlandi, Eni Pulcinelli (1996). A linguagem e seu funcionamento: as formas do discurso. (4. ed.). Campinas: Pontes, 1996.

Orlandi, Eni Pulcinelli (1999). Análise do discurso: princípios e procedimentos. Pontes. Campinas: Pontes, 1999.

Orlandi, Eni Pulcinelli (2002). Análise de discurso. Princípios e procedimentos. 4. ed. Campinas: Pontes, 2002.

Orlandi, Eni Pulcinelli (2008). Discurso e texto: formulação e circulação de sentidos. Campinas: Pontes, 2008.

Orlandi, Eni Pulcinelli (2009). O que é linguística. São Paulo: Brasiliense, 2009.

Orlandi, Eni Pulcinelli (2020). Análise do discurso: princípios e procedimentos. (13. ed.). Campinas: Pontes, 2020.

Pêcheux, Michel (1969). Análise automática do discurso (AAD-69). // Gadet, F.; Hak, T. (eds.). Por uma análise automática do discurso: uma introdução à obra de Michel Pêcheux. Campinas: Editora da Unicamp, 2014. 59-158.

Pêcheux, Michel (1998). Sobre os contextos epistemológicos da Análise de Discurso. Tradução Ana Maria Dischinger e Heloisa Monteiro Rosário. // Cadernos de Tradução 1. Porto Alegre: Editora da UFRGS, 1998.

Pêcheux, Michel (2014). Semântica e discurso: uma crítica a afirmação do óbvio. (5. ed.). Tradução Eni Puccinelli Orlandi, Lourenço Chacon Jurado Filho, Manoel Luiz Gonçalves Corrêa e Silvana Mabel Serrani. Campinas: Editora da Unicamp, 2014..

Pêcheux, Michel (2015) O discurso: estrutura ou acontecimento. (7. ed.). Tradução Eni P. Orlandi. Campinas: Pontes, 2015.

Radford, Gary P. (2003). Trapped in our own discursive formations: toward an archaeology of library and information science. // The Library Quarterly 73:1 (2003) 1-18.

Radford, Gary P.; Radford, Marie L. (2005). Structuralism, Post-Structuralism, and the Library: de Saussure and Foucault. // Journal of Documentation. 61:1 (2005) 60-78. https://doi.org/10.1108/00220410510578014

Reis, Daniela Majorie Akama dos (2019). A leitura documentária de bibliotecários jurídicos: um estudo realizado a partir de aspectos da semiose e teoria da inferência observados na estrutura textual de doutrina. [tese de doutorado]. Marília: Universidade Estadual Paulista, 2019. http://hdl.handle.net/11449/181849

Santaella, Lucía (2005). Matrizes da linguagem e pensamento: sorora, visual e verbal: aplicações na hipermídia. 3. ed. São Paulo: Iluninuras, 2005.

Silva, Eduardo Graziosi; Boccato, Vera Regina Cassari (2012). Avaliação do uso de catálogos coletivos de bibliotecas universitárias pela perspectiva sociocognitiva do usuário. // TransInformação. 24:1 (2012) 05-18.

Simon, Robert I. (2014). Homens maus fazem o que homens bons sonham: um psiquiatra forense ilumina o lado obscuro do comportamento humano. Porto Alegre: Artmed, 2014.

Stolz, Sheila (2008). A ordem do discurso e suas relações com o poder: vertigem e quebra de certezas. // JURIS: Revista da Faculdade de Direito. 13 (jan./dez. 2008) 159–176.

Enviado: 2024-04-20. Segunda versión: 2024-05-16.
Aceptado: 2024-06-05.

First steps towards a platform for the analysis of civil law documentary heritage

Primeros pasos hacia una plataforma para el análisis del patrimonio documental de derecho civil

Hala NEJI (1), Javier NOGUERAS-ISO (1),
Francisco Javier GARCÍA-MARCO (2), Carmen BAYOD LÓPEZ (2)

(1) Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, C/ Mariano Esquillor SN, 50018 Zaragoza, España, {hala.neji,jnog}@unizar.es. (2) Instituto de Patrimonio y Humanidades (IPH), Universidad de Zaragoza, C/ Pedro Cerbuna 12, 50009 Zaragoza, España, {jgarcia,cbayod}@unizar.es

Resumen

El patrimonio documental sobre derecho civil es un importante activo cuyo estudio permite conocer el contexto político, social y cultural de la época a la que se refieren los documentos históricos. Este trabajo presenta el diseño de un prototipo de plataforma para apoyar a los investigadores en el análisis del patrimonio documental sobre derecho civil. Nuestra plataforma consiste en crear una versión en línea de estos materiales, haciéndolos más accesibles. La plataforma proporciona asistentes automáticos para la transcripción, traducción y extracción de elementos de información específicos asociados a conceptos de derecho civil (voces) como citas a fuentes externas y entidades con nombre (lugares, personas y organizaciones) para identificar mejor su contexto. La viabilidad de esta plataforma se ha puesto a prueba con el tratamiento de una obra doctrinal escrita por Miguel del Molino, un conocido experto en derecho civil del siglo XV en el reino de Aragón.

Palabras clave: Derecho civil. Patrimonio documental. Procesamiento de textos. Reconocimiento de entidades nombradas. Aprendizaje automático. Molino, Miguel del.

Abstract

The documentary heritage about civil law is an important asset whose study allows us to learn about the political, social and cultural context of the period referred in historical documents. This paper presents the design of a prototype platform to sup-port researchers in the analysis of civil law docu-mentary heritage. Our platform involves creating an online version of these materials, making them more accessible. The platform provides automatic assistants for the transcription, translation and extraction of specific information items associated to civil law concepts (voices) such as citations to external sources and named entities (locations, persons, and organizations) to identify better their context. The feasibility of this platform has been tested with the processing of a doctrinal work writ-ten by Miguel del Molino, a well-known civil law expert in XV century in the Aragon kingdom.

Keywords: Civil law. Documentary heritage. Text processing. Named entity recognition. Deep learning. Miguel del.

1. Introduction

Civil law regulates the civil or private relations of persons: it deals with the civil status of persons, their family rights and duties, property and other real rights over things, the regime of obligations and contracts, and successions and inheritances. Although in Europe each country has its own civil law code at the state level, this law dates back to past times and has evolved until present times. In many administrative areas this law has its origin in historical kingdoms that regulated the conditions of settlement of their citizens. The books that publish these historical codes of civil law and the doctrinal works that document their interpretation make up an asset of cultural interest whose study allows us to learn about the political, social, and cultural context of this historical period.

Nowadays there are numerous initiatives (e.g. Europeana) that have promoted the digitization of documentary heritage, including that related to civil law. However, simple digitization is not enough. This work is a first step in a research project that aims to promote the study of works of special interest in the field of the history of civil law (specially, civil law authors of XVI, XVII and XVIII centuries) accompanied with the development of the necessary technology to have an online hypertext edition of the works that includes the digital facsimile, a critical edition, a translation and all the complements that are deemed relevant (e.g., legal codes cited and notes or glosses to the text both legal and historical-philological or bibliographical).

The objective of this work is to present a first prototype of the platform that will provide support for the analysis of civil law documentary heritage. Figure 1 presents a use case diagram which outlines various functionalities of the envisioned web

platform. Apart from the typical tools for searching, browsing and visualization available for the general public, this platform integrates advanced tools to support transcription, translation, and legal analysis of the books ingested by experts in civil law. With respect to transcription and translation, we aim to customize existent OCR and automated translation algorithms. Related to the legal analysis, we aim to integrate text mining tools able to extract context information from the text defining how civil law concepts should be interpreted. This context consists of citations to regulations and named entities such as locations or authorities (persons and organisations).

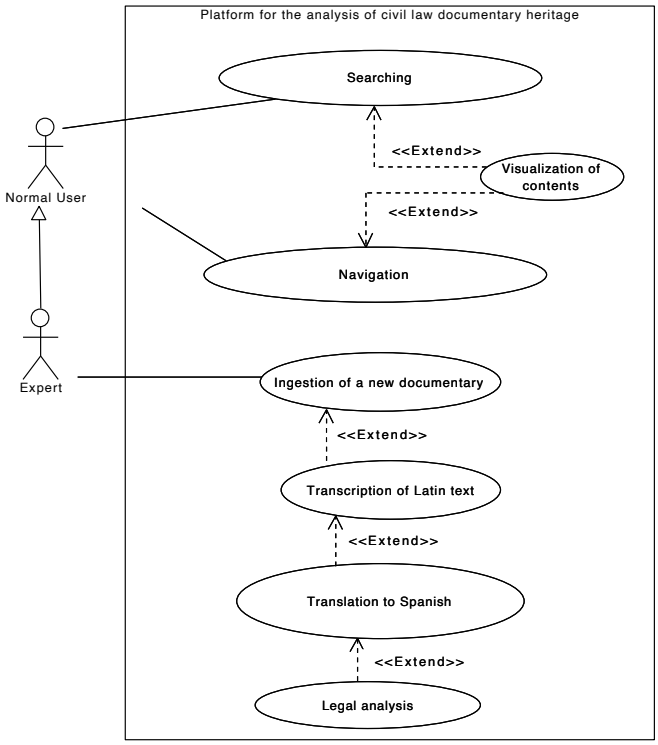

Figure 1. Use case diagram

The rest of this work is organized as follows. Section 2 reviews the state-of-the-art in documentary heritage analysis. Section 3 describes the design of the platform. Then, section 4 presents some experimental results of the platform for the analysis of a work of Miguel del Molino, a well-known civil law expert in XV century in the Aragon kingdom (currently an autonomous community in Spain). Finally, section 5 provides some concluding remarks.

2. State of the art

This platform presents important challenges regarding transcription, translation and named entity recognition.

With respect to the *transcription* task, it must be noted that the images from the digitization of historical documents present various levels of degradation due to the use of handmade fonts, ink stains on the paper or the noise generated during the digitization process (interference from double-sided printed text, deviations, blurring in double-page scans, etc.) (Gupta et al., 2015). All of this makes optical character recognition (OCR) difficult and motivates research into the pre-processing of images to eliminate noise (Neji et al., 2024), or the training of character recognition models specialized in Gothic and round letters using machine learning based on neural networks (Lacasta et al., 2022; Kodym et al., 2021).

It is also worth noting the existence of the Pero-OCR tool, which is semi-supervised machine learning method for automatic handwritten and printed text transcription (Kišš et al., 2023). It employs a SoftCTC (Connectionist Temporal Classification) loss function that allows to manage complex transcription scenarios.

In addition, automated *translation* from Latin to Spanish also represents an important challenge. Although there are currently numerous online translators (Google Translator, Yandex, DeepL, Translateking, imTranslator or Translateking, among others) and some configurable open source tools (for example, NVIDIA NeMo or OpenNMT), the translation of medieval Latin is not very advanced due to the scarcity of parallel Latin and Spanish corpora in different domains (Tiedemann et al., 2012).

There are incipient works that exploit deep neural networks based on transformer-type architectures (Transformer) for this type of problems. For instance, Martínez García and García Tejedor (2020) developed an advanced Neural Machine Translation system for Latin-Spanish, aiming to make historical texts more accessible. Using Transformer-based models trained on Bible and Saint Augustine corpora, the study explores domain adaptation challenges and emphasizes the significance of sufficient data for accurate translations, especially for low-resourced languages like Latin, but for the moment each translation work in this context requires the preparation of a personalized training corpus. Fischer et al. (2022) also crafted a dedicated Latin-German Neural Machine Translation (NMT) system for 16th-century letter translation. Their meticulous data collection and NMT model development led to superior translations for short to medium sentences, outperforming GoogleTranslate. While centered on Swiss reformer Heinrich Bullinger's correspondence, their work offers broader utility for translating texts from the era.

Thirdly, to facilitate the *legal analysis* of the works, it is relevant to adapt text mining techniques that allow the recognition of named entities such as place names, person names and organisation names to contextualize the voices included in the doctrinal works of civil law. Identifying people, places, and other historical entities is an essential task in automatic understanding of historical documents (Aljalbout and Falquet, 2017). Named entity recognition and classification (NER for short) is very often the first step of entity linking, which can support the cross-linking of multilingual and heterogeneous heritage collections based on authority files and knowledge bases and can greatly support the search and exploration of historical documents. Nowadays there are several approaches making profit of machine learning methods for NER (Li et al. 2020). For instance, Erdmann et al. (2016) presented a CRF-based model with handcrafted features for Latin historical texts and motivated the choice of Part-of-Speech (POS) tagger by the fact that this NLP tool leverages the highly informative morphological complexity of Latin. Hubkova at al. (2019) also proposed a BiLSTM-based model by applied a character-based CNN to encode the different spellings of words. Nonetheless, it must be noted also the difficulties for applying NER to historical documents and the consequent degradation of performance metrics (Rodriguez et al., 2012; van Strien et al., 2020). In such a context it is essential to count on annotated corpus and benchmarks. For instance, Hubkova et al. (2019) curated and annotated a corpus from scanned Czech historical newspapers. In the same line, Hamdi et al. (2019) delineated a German gold standard for NER within the domain of historical biodiversity literature.

3. Design of the platform

Figure 2 presents the architectural design of our proposed web platform. The architecture contains three main components: database, content management, and search/visualization.

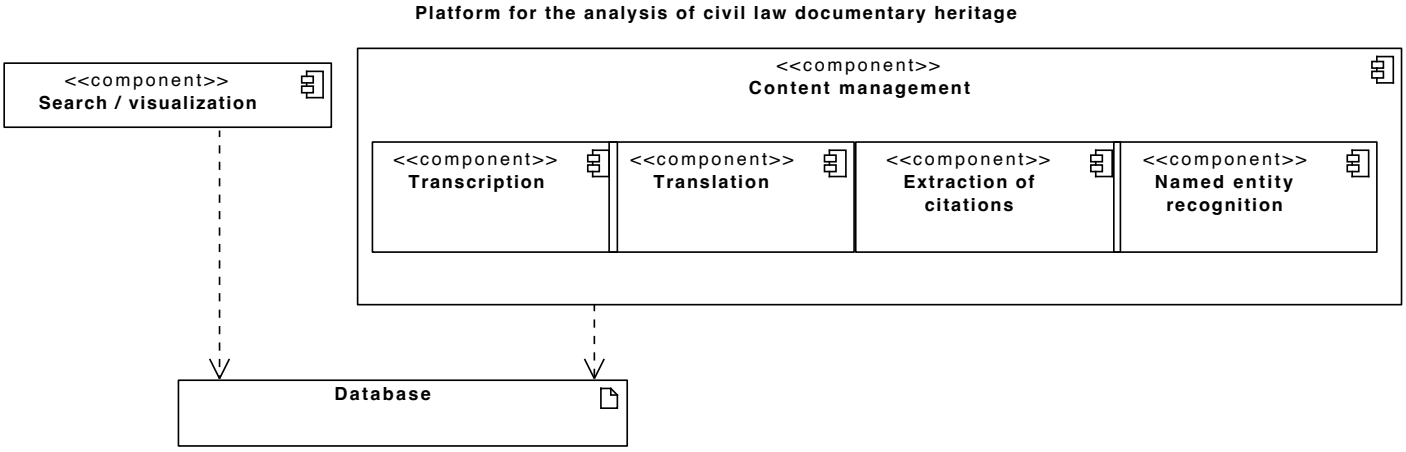

Figure 2. Architecture of our platform

Firstly, the *database* provides a semantic repository (RDF triplestore) for representing the structural contents of a civil law book. The type of books that we aim to analyse in this platform consist of a set of civil law concepts (voices) with its associated juridical interpretation. For instance, the left side of Figure 3 shows an example of the table of contents in the work *Repertorium Fororum et Observantiarum Regni Aragonum: una pluribus cum determinationibus consilii iustitiae Aragonum practicis atquae cautelis eisdem fideliter annexis* written by Miguel del Molino in an edition published in 1585. On the right side of the figure, the initial page containing the *Adulterium* concept is displayed. Considering this, Figure 4 shows the main classes and properties proposed for the RDF triplestore. Whenever possible, we have reused terms from well-known vocabularies such as the DCMI Metadata Terms (terms with *dct* prefix), the DCMI Type vocabulary (terms with *dctype* prefix), the FOAF vocabulary (terms with *foaf* prefix), the Simple Knowledge Organization System (SKOS, terms with *skos* prefix) and the Core Location Vocabulary (terms with *locn* prefix).

In addition, we have defined new classes and properties in our *Civil* vocabulary to define better the peculiarities of a book of civil law (*civil:Book*) as an extension of a printed text resource (*dctype:Text*). It must be noted also that we have defined a particular class for representing the civil law concepts (*civil:Concept*) discussed in these books. This civil law concept is an extension of the

typical concept (*skos:Concept*) in a knowledge organisation system. Apart from a preferred label, it is also annotated with properties containing links to digitized page images (*dct:source*), the original transcribed text associated to the concept (*civil:transcription*), its translation into a modern language (*civil:translation*), and any type of reference (*dct:references*) to bibliographic resources, locations, persons or organizations.

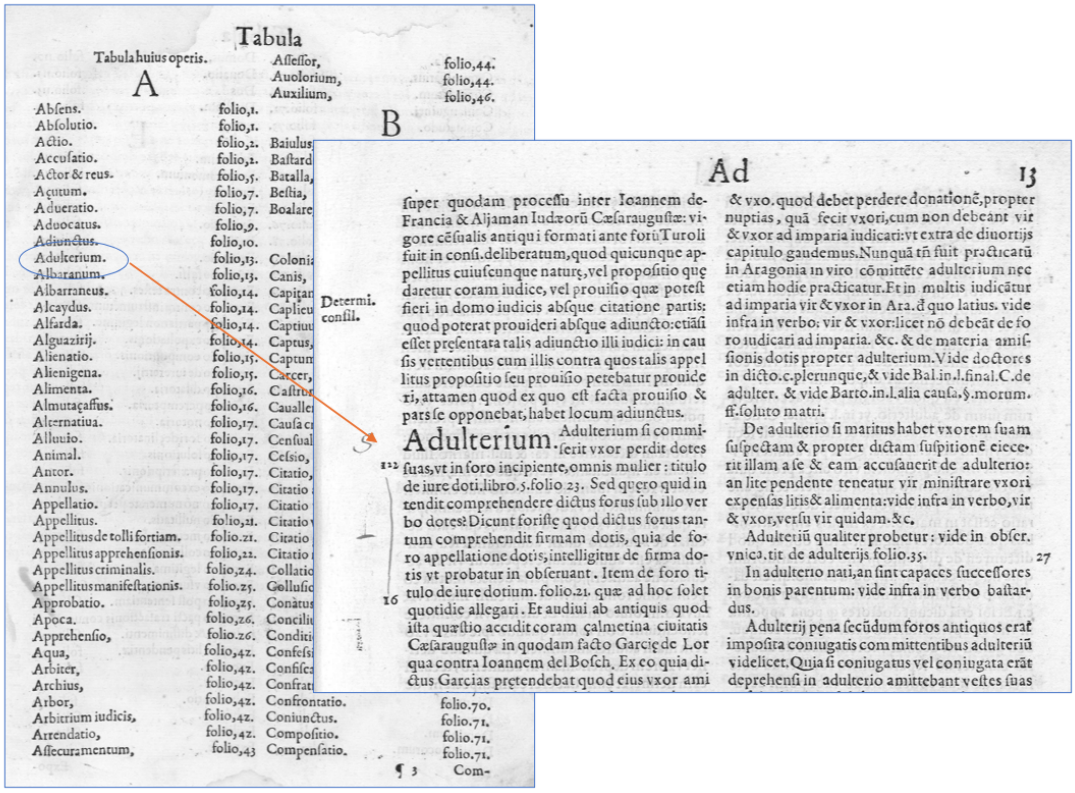

Figure 3. An example of pages from Miguel del Molino's book

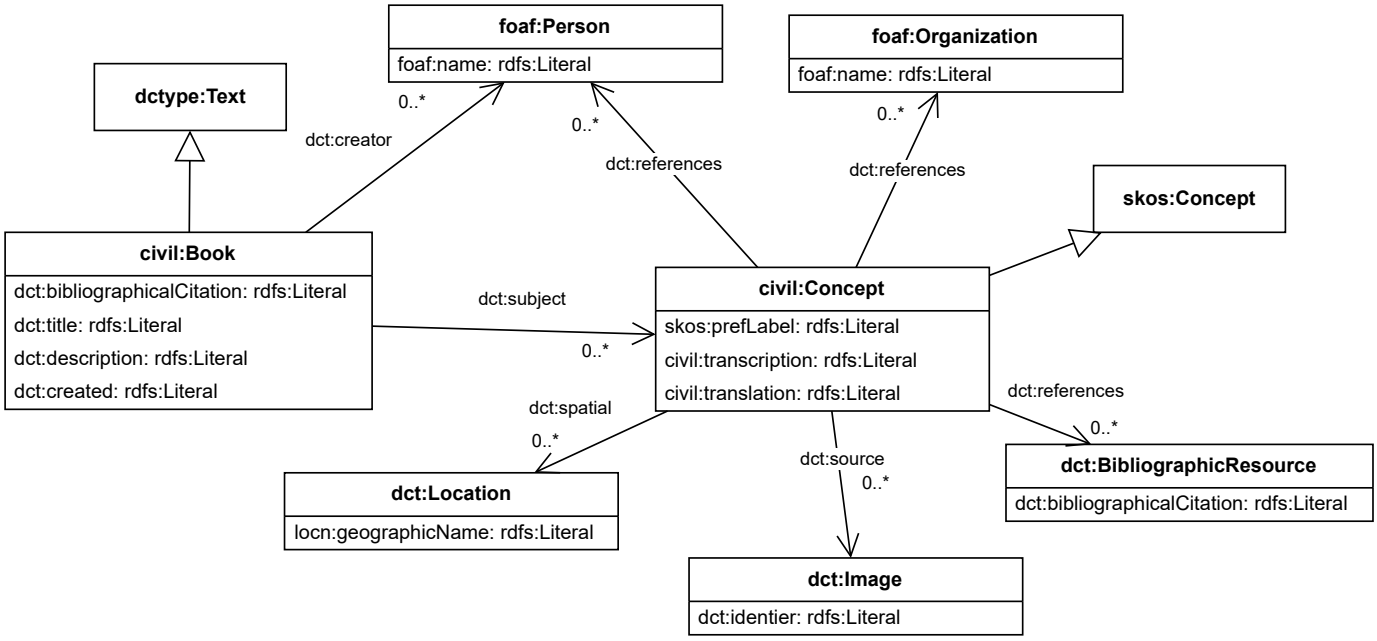

Figure 4. Conceptual model (UML class diagram) of the semantic repository

Secondly, the *content management* component cordinates a set of automated tasks to facilitate the processing and analysis of documents. Figure 5 shows an activity diagram representing the processing workflow applied to a new book ingested in the platform and the inputs/outputs that are generated.

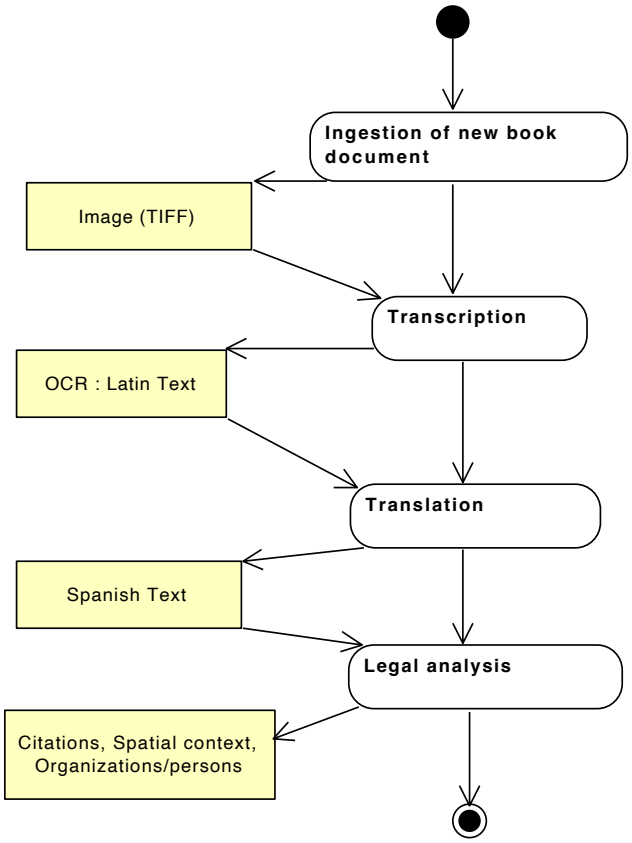

Figure 5. Workflow for the processing of civil law documents

On the one hand, the transcription component aims to convert the digitized images of book pages with printed text into machine-readable textual format. For this step, we have used Pero-OCR (Kišš et al. , 2023), a state-of-the-art optical character recognition (OCR) tool. The transcription process, apart from obtaining the OCR of every page in the book, also encompasses the division of text for the different civil law concepts. On the other hand, the translation component converts the original latin text into a modern language (e.g. Spanish). As a first approach for this translation task we have used the Google Translator API to translate the Latin text linked to each civil law concept into Spanish. Moreover, we have also components for citation extraction and named entity recognition. They contribute to the semantic understanding and analysis of document content, laying the groundwork for advanced information retrieval and manipulation. For the identification of citations, we have implemented the detection of some string patterns followed by the external citations. In the case of named entity recognition, we have integrated the use of *spacy*, an open-source Python library for Natural Language Processing the includes pre-trained models for the identification of locations and authorities in Spanish texts.

Last, the *search/visualization* component facilitates the interaction between the users and our platform. There are two types of users: normal and expert users.

Normal users are researchers or general public interested in civil law. For these users the platform offers a direct access to a homepage displaying the list of books (Figure 6) and enabling the export of book contents in RDF format. In addition, the homepage allows the filtered search on books and specific contents of a book through facets for Concepts, Locations, Persons, and Organizations where the user can either type text or select a value from a list. The search results are presented in a web page which displays a list of all concepts that meet the search criteria specified on the search page. For example, Figure 7 presents a search result after filtering the concept name *Albarraneus*.

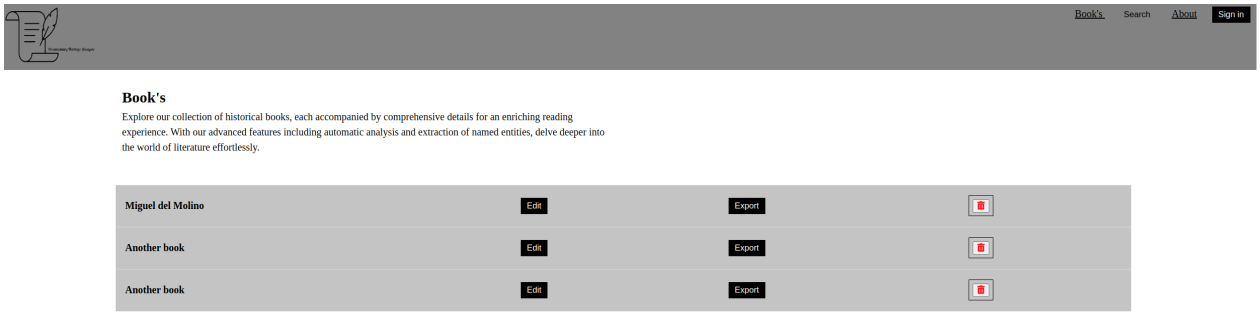

Figure 6. Homepage displaying the list of books

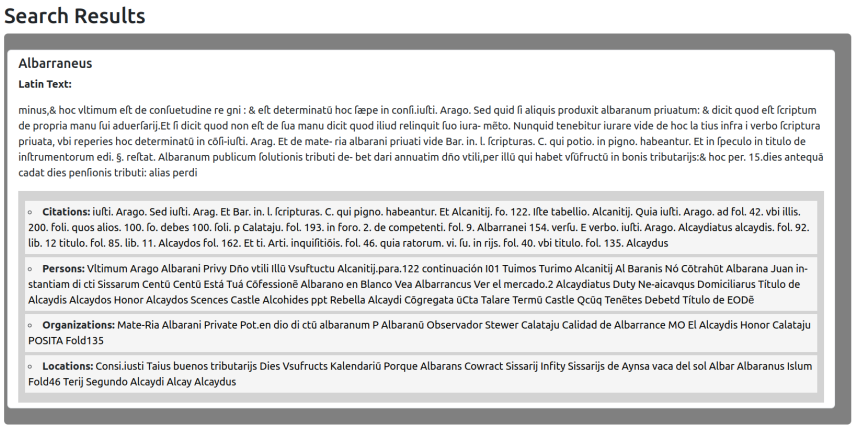

Figure 7. Search Results web page

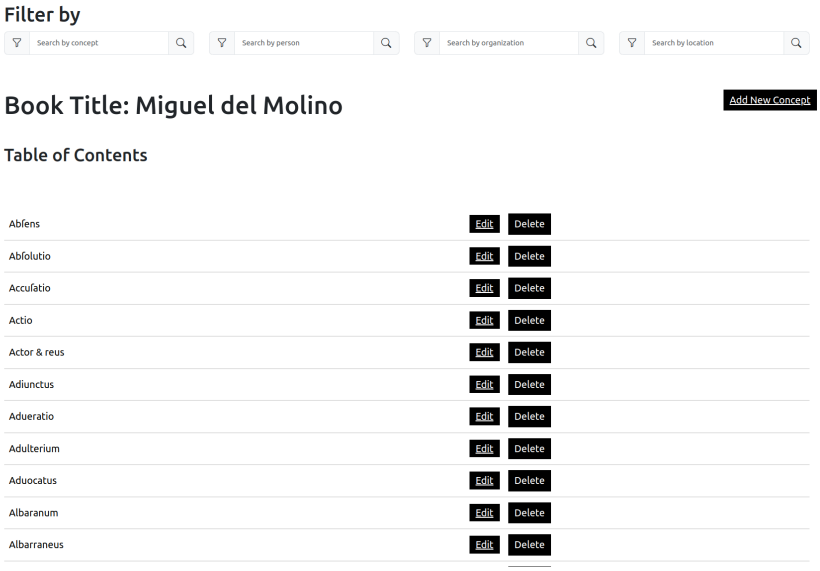

Figure 8. Management of the list of concepts in a book

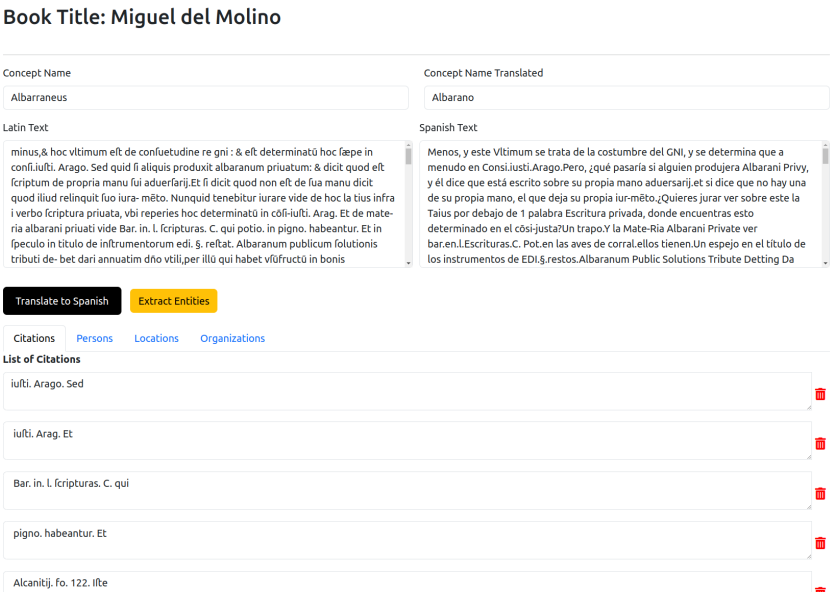

Figure 9. Edit concept web page

The experts are the authorized users that can add, modify or delete the books and their associated concepts (Figure 8, in previous page). In addition, they have access to the advanced functionalities for automatic transcription, translation, and information extraction in the web page for editing concepts (Figure 9, in previous page).

4. Experimental results

The proposed platform for the analysis of documentary heritage, whose design was presented in Section 3, has been implemented using the Django framework for web development. In addition, the components for content management have been implemented in Python integrating different libraries for OCR (*Pero-OCR*), and named entity recognition (*spacy*).

Moreover the feasibility of the implementation has been tested with the processing of a doctrinal civil law book written by Miguel del Molino and printed in 1585: *Repertorium Fororum et Observantiarum Regni Aragonum: una pluribus cum determinationibus consilii iustitiae Aragonum practicis atquae cautelis eisdem fideliter annexis*, available at

https://derechoaragones.aragon.es/es/consulta/registro.do?id=600036

In particular, we were interested in the performance of the automated assitants for transcription, translation and information extraction.

With respect to the transcription, we compared the results of Pero-OCR with the results obtained after applying Tesseract, a widely used open-source OCR tool (https://github.com/tesseract-ocr/tesseract). The experience showed that Pero-OCR is more efficient. For instance, Figure 10 depicts the transcription of a sample text with both Tesseract and Pero-OCR. There are letters that were not recognized by Tesseract , such as the letter "s" (ſ contained in *abſens*), which in the Latin language is very similar to "f". In contrast, it was correctly detected by Pero-OCR.

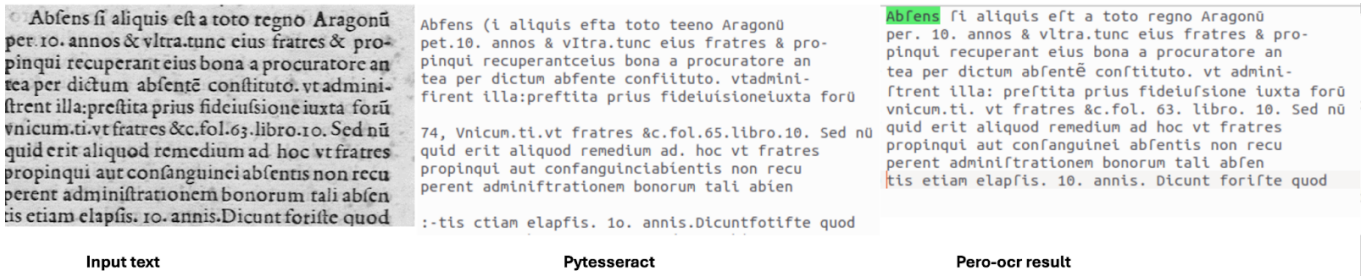

Figure 10. Comparison between Tesseract and Pero-OCR to generate a text

With respect to the translation, we compared qualitatively the results obtained after invoking the API of several online translators: Google Translator, Yandex, DeepL, Translateking, imTranslator and Translateking. After comparing the results obtained with some sample texts directly generated by Pero-OCR, we considered that the Google Translator API was providing the best results.

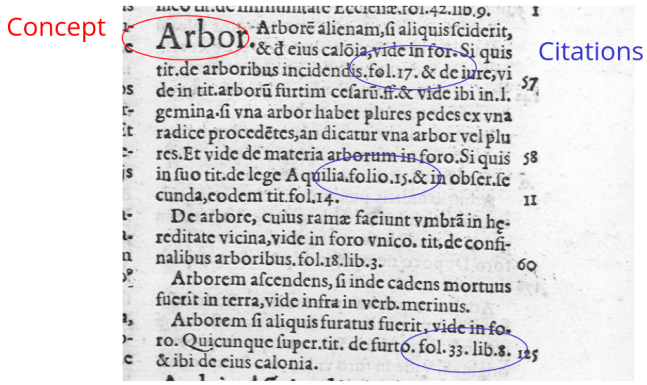

Figure 11. Example of concept and citations

Regarding the extraction of citations, we concluded that this processing task must be customized to the special features of each book. Figure 11 shows part of the text associated to the concept “Arbor” and we highlighted some external citations to *“Fori Regni Aragonum”* (https://derechoaragones.aragon.es/es/consulta/registro.do?id=600013), a collective work printed in 1496 that compiles the official regulation books about civil law by the end of the XV century. Even in this small example we can identify some patterns to cite a book in this collective work. For instance, from the text:

> […] vide in foro. Quicunque super.tit. de furto.fol.33.lib.8.

we can derive the following citation pattern:

> […] vide in foro. <beginning words of a paragraph within a title>.tit. de <title name>.fol.<page number>.lib.<book number>.

Figure 12 shows an excerpt of the cited page of the external book.

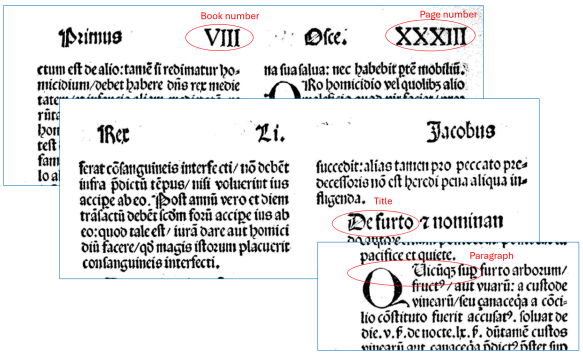

Figure 12. Example of cited title and paragraph in page 33 of book 8.

About the direct applicability of the *spacy* library for named entity recognition (NER), the quality of results is not very satisfactory because the translation into Spanish contains frequent mistakes, derived in turn by some errors found in the Latin transcription. Although the results of NER could be improved once a better translation into Spanish is provided, we believe that the incorporation of knowledge bases such gazetteers of historical place names or authority files would help to identify better these named entities.

5. Conclusions

This paper has presented a first prototype for the development of a web platform to facilitate the analysis of civil law documentary heritage. The design of the platform has been customized to the specific needs of the experts in this domain. First, we have designed a conceptual framework to represent the information required for a deep analysis of civil law contents. In addition, we have integrated open-source tools to assist in tasks such as transcription, translation, and information extraction, which usually require high human resources if performed completely manually.

Although the results obtained by automated tasks could be clearly improved, they provide an appropriate input in each step that can be revised by experts to ameliorate the performance. For instance, if the Latin transcription of a text is revised by an expert in the platform, the automated translation will provide better results.

As future work, we will continue with the development of the platform and the testing of more alternatives for assisting in the automated tasks of transcription, translation, and information extraction. This further development will be accompanied with a detailed analysis of the results obtained with the ingestion of a more extensive corpus of civil law books in the platform. In addition, a version manager to control variants among different manuscripts, editions and transcriptions is being considered. Finally, a system for collaborative annotation and linking of the documents to support advanced academic research is also envisaged (García-Marco, 2020).

Acknowledgements

This paper is part of the R&D projects T59_23R and S15_23R supported by the Aragon Regional Government in cooperation with the “Cátedra de Derecho civil y foral de Aragón”.

References

Aljalbout, S.; Falquet, G. (2017). Un modèle pour la représentation des connaissances temporelles dans les documents historiques: Applications sur les manuscrits de F. Saussure. // Proc. 28es Journées francophones d'Ingénierie des Connaissances (IC 2017): Caen, France, July 2017.

Erdmann, A.; Brown, C.; Joseph, B.D; Janse, M.; Ajaka, P.; Elsner, M.; de Marneffe, M. (2016). Challenges and solutions for Latin named entity recognition. // COLING 2016: 26th International Conference on Computational Linguistics. Association for Computational Linguistics. 85–93.

Fischer, L.; Scheurer, P.; Schwitter, R.; Volk, M. (2022). Machine translation of 16th century letters from Latin to German. // Proceedings of 2nd Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) at LREC-2022, Marseille.

García Marco, Francisco Javier. Knowledge Organization in Historical Information Systems Revisited: Changes in Society, Technology and Expectations 25 Years Later. // Knowledge Organization at the Interface. Proceedings of the Sixteenth International ISKO Conference 6-8 July 2020 Aalborg, Denkmark. Würzburg: Ergon-Verlag GmbH, 2020. 474-478.

Gupta, A.; Gutierrez-Osuna, R.; Christy, M.; Capitanu, B.; Auvil, L.; Grumbach, L.; Furuta, R.; Mandell, L. (2015). Automatic Assessment of OCR Quality in Historical Documents. // Proc. of 29th AAAI Conference on Artificial Intelligence. 1735-1741.

Hamdi, A.; Jean-Caurant, A.; Sidere, N.; Coustaty, M.; Doucet, A. (2019). An analysis of the performance of named entity recognition over ocred documents. // 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE. 333–334.

Hubkova, H. (2019). Named-entity recognition in Czech historical texts: Using a CNN-BiLSTM neural network model. Ph.D. thesis.

Kišš, M.; Hradiš, M.; Beneš, K.; Buchal, P.; Kula, M. (2023). SoftCTC: semi-supervised learning for text recognition using soft pseudo-labels. // International Journal on Document Analysis and Recognition (IJDAR). 2, 1-17.

Kodym, O.; Hradiš, M. (2021). Page layout analysis system for unconstrained historic documents. // Proc. of 16th International Conference on Document Analysis and Recognition–ICDAR 2021: Lausanne, Switzerland, September 5–10, 2021. Part II, 492-506).

Lacasta, J.; Nogueras-Iso, J.; Zarazaga-Soria, F.J.; Pedraza-Gracia, M.J. (2022). Tracing the origins of incunabula through the automatic identification of fonts in digitised documents. // Multimedia Tools and Applications. 81:28, 40977-40991.

Li, J.; Sun, A.; Han, J.; Li, C. (2020). A survey on deep learning for named entity recognition. // IEEE Transactions on Knowledge and Data Engineering. 34:1, 50-70.

Martínez Garcia, E; García Tejedor, Á. (2020). Latin-Spanish Neural Machine Translation: From the Bible to Saint Augustine. // Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages: Marseille, France. European Language Resources Association (ELRA). 94–99.

Neji, H.; Ben Halima, M.; Nogueras-Iso, J.; Hamdani, T.M.; Lacasta, J.; Chabchoub, H.; Alimi, A.M. (2024). Doc-Attentive-GAN: attentive GAN for historical document denoising. // Multimedia Tools and Applications. 83, 55509–55525.

Rodriquez, K.J.; Bryant, M.; Blanke, T.; Luszczynska, M. (2012). Comparison of named entity recognition tools for raw OCR text. // 11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012. Scientific series of the OGAI. 5, 410–414.

Tiedemann, J. (2012). Parallel data, tools and interfaces in Opus. // Calzolari, Nicoletta (Conference Chair); et al., (eds). Proc. of 8th Int. Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA).

van Strien, D.; Beelen, K.; Coll Ardanuy, M.; Hosseini, K.; McGillivray, B.; Colavizza, G. (2020). Assessing the impact of OCR quality on downstream NLP tasks. // Proceedings of the 12th International Conference on Agents and Artificial Intelligence. 1, 484-496. https://doi.org/10.5220/0009169004840496

Enviado: 2024-06-07. Aceptado: 2024-06-11.

O conteúdo é rei? Fatores determinantes numa estratégia de SEO para media digitais

¿El contenido es el rey? Factores clave de una estrategia SEO para medios digitales

Is content king? Key factors of an SEO strategy for digital media

Branco DI FÁTIMA (1), Diogo GIL (2)

(1) LabCom, Universidade da Beira Interior (UBI), Rua Marquês D'Ávila e Bolama, 6201-001, Covilhã, Portugal, brancodifatima@labcom.ubi.pt. (2) Escola Superior de Comunicação Social (ESCS), Campus de Benfica do IPL, 1549-014, Lisboa, Portugal, diogoaggil@gmail.com

Resumen

Se analiza la influencia de los principales factores SEO en el éxito de una estrategia de optimización para medios digitales. El punto de partida es la conocida expresión "el contenido es el rey", pronunciada por Bill Gates en 1996. La muestra se compone de 12 websites portugueses de educación financiera, considerados referencias nacionales en este campo de conocimiento. Los datos se extrajeron utilizando la versión de pago de Ubersuggest, y corresponden a un período de un año. Los resultados sugieren que el contenido no es el factor más importante en una estrategia SEO. Sin embargo, factores como los *backlinks*, la autoridad del dominio y las redes sociales están directamente relacionados con el tráfico orgánico.

Palabras clave: SEO. Motores de búsqueda. Google. Métodos digitales. Ubersuggest. Educación financiera. Brasil.

Abstract

The key SEO factors and their impact on the effectiveness of a digital media optimization strategy are analyzed. It begins with a reference to Bill Gates' famous statement, "content is king," dating back to 1996. The study is based on a sample consisting of 12 prominent Portuguese financial education websites, recognized as national benchmarks in their field. Data collection was facilitated using the premium version of Ubersuggest, covering a comprehensive timeframe of one year. The findings challenge the traditional belief that content is the most crucial aspect of an SEO strategy. Instead, they highlight the importance of factors such as backlinks, domain authority, and social media presence, which show a direct correlation with organic web traffic.

Keywords: SEO. Search engines. Google. Digital methods. Ubersuggest. Financial education. Brazil.

1. Introdução

Uma estratégia de *Search Engine Optimization* (SEO) tem o objetivo de conseguir o melhor posicionamento nos resultados orgânicos de um motor de pesquisa, baseando-se na relevância do website para os utilizadores, sem envolver custos diretos com publicidade (Ortega, 2020). Mas, porque isto é tão importante?

De acordo com o Advanced Web Ranking (2021), as quatro primeiras posições orgânicas no resultado das pesquisas têm cerca de 60,0 % dos cliques. Quanto mais afastada uma página web estiver da primeira posição, menor é a probabilidade de ser visitada. Por causa disso, nos últimos anos, tem crescido a importância dos profissionais especializados em SEO (Escandell-Poveda, Papí-Gálvez e Iglesias-García, 2023).

Este artigo examina a influência dos principais fatores de SEO em uma estratégia bem conseguida de otimização para media digitais. O ponto de partida é a já conhecida expressão "o conteúdo é rei", dita por Bill Gates, em 1996. O objetivo é responder se, para a Google e o seu algoritmo de *ranking*, o *PageRank*, o conteúdo é mesmo o fator mais importante numa estratégia de SEO.

A amostra é constituída por 12 websites portugueses de literacia financeira, considerados de referência nacional nesse campo do saber. Os dados foram extraídos com a versão paga da Ubersuggest, e correspondem ao intervalo temporal de um ano.

É verdade que a classificação dos motores de pesquisa baseia-se na popularidade dos websites (Pedrosa e Morais, 2021). No entanto, a popularidade é calculada por métricas que indicam a relevância de um tema, uma notícia, um fenómeno, etc. Por esse caminho, os motores de pesquisa estão a estabelecer os paradigmas sobre o que é prestigiado ou não na sociedade hiperconectada.

2. Marco teórico

Os motores de pesquisa são websites que permitem identificar e recuperar informação de outras páginas web. Contudo, a história dessas ferramentas é mais antiga que a própria web (Ippolita, 2013; Seymour, Frantsvog e Kumar, 2011).

Desde as primeiras iniciativas nesse sentido, com o W3Catalog e o Aliweb, no final de 1993, até ao monopólio da Google, as tecnologias digitais assumiram um papel preponderante na vida em sociedade (Castells, 2012). Cada vez mais, os motores de pesquisa são uma ponte entre o utilizador de internet e informações muito diversas, sobre política, agricultura, meteorologia, saúde ou finanças (Sanchez-Cuadrado e Morato, 2023).

O *Global Overview Report*, do Hootsuite (2021), mostra que 60,0% da população mundial têm acesso à internet, com a participação expressiva dos povos da Europa (93,0 %) e da América do Sul (72,0 %). Esses indivíduos tomam decisões baseadas em informações encontradas online – um fenómeno acentuado pela pandemia de Covid-19 (Romero-Sánchez e Barrios-Hernández, 2023; Feldmann *et al.*, 2021). Logo, é possível assumir que a importância do SEO se deve a um conjunto de fatores, como o aumento de utilizadores da internet, o crescimento no número de websites e a complexificação dos motores de pesquisa (Lopezosa, Guallar e Santos-Hermosa, 2022; Faustino, 2019).

Porém, os resultados dos motores de pesquisa nem sempre foram tão vastos e precisos. A verdadeira mudança nessa área – e ascensão do SEO – está no surgimento da Google e do seu algoritmo: o *PageRank* (Duong, 2019). O ponto de viragem foi a combinação de inúmeros fatores – que podem chegar a 200 – para devolver os resultados pretendidos (Dean, 2023).

Como os utilizadores pesquisam assuntos diferentes e têm um histórico de navegação igualmente diverso, o objetivo de uma página web nunca será aparecer em primeiro lugar em todas as pesquisas. De qualquer forma, existe uma regra mais ou menos acordada: quanto mais afastado um website estiver das primeiras posições, menor será a probabilidade de ser visitado (Advanced Web Ranking, 2021).

O algoritmo de qualquer motor de pesquisa tem primeiro de fazer o rastreamento dos websites, depois a sua indexação e, por fim, elaborar o *ranking* de qual ficará em primeiro, em segundo e por aí em diante (Jain, 2013). Assim, o que é valorizado pelos algoritmos torna-se também o essencial de uma estratégia de otimização das páginas web (Finch, 2019).

Nesse sentido, os fatores de SEO são normalmente organizados em dois campos temáticos (Pedrosa e Morais, 2021; Gudivada *et al.*, 2015; Sebring, 2019; Duong, 2019):

- O SEO *on-page* corresponde ao conjunto das técnicas aplicado dentro do website, ou seja, que estão sob o controlo do dono da página web. Segundo Palanisamy e Liu (2018), agrupa fatores utilizados para alcançar uma classificação melhor no resultado. As técnicas *on-page* podem fazer parte da própria codificação do website, como a introdução de palavras-chave, títulos, links internos, descrições de imagens, *layout* responsivo, entre outros.
- O SEO *off-page* trabalha com técnicas fora do website, ou seja, que fogem ao controlo absoluto do dono da página web. De acordo com Gudivada *et al.* (2015), são os fatores associados às ligações externas e que direcionam ao website, também chamadas de *backlinks*. Contribuem para aumentar a reputação, ganhar autoridade e melhorar a classificação da página nos resultados de pesquisa. Assim, as redes sociais são um elemento importante para o tráfego web.

2.1. Fatores SEO mais significativos

Frente aos inúmeros fatores *on-page* e *off-page*, especialistas têm indicado os mais significativos em uma estratégia bem conseguida de otimização para media digitais (Drivas *et al.*, 2020; Patel, 2021; Faustino, 2019).

Um dos fatores mais relevantes são as ligações internas entre as páginas web. O objetivo é facilitar o rastreamento e a indexação de componentes individuais, promovendo a partilha de autoridade entre as várias páginas (Faustino, 2019). O *PageRank* valoriza as estruturas hipertextuais complexas, enquanto ligações danificadas podem gerar um efeito inverso (Yussuf, 2020; Ziakis *et al.*, 2019).

A autoridade corresponde à relevância de um website quando comparado aos demais. O cálculo é impactado pelos chamados *backlinks*, ou seja, os links externos que apontam para o website (Patel, 2021). Os *backlinks* estão entre os fatores mais importantes para a Google, representando um voto de relevância do conteúdo (Lopezosa *et al.*, 2019).

O *PageRank* também categoriza os *alt text* – conteúdos textuais embutidos nas imagens (Patil-Swati *et al.*, 2013). O *alt text* pode ajudar na utilização de programas de leitura por deficientes visuais, como texto alternativo, uma vez que a imagem é descrita (Chiarella, Yarbrough e Jackson, 2020).

O algoritmo da Google determina quantas vezes uma expressão ocorre em relação ao restante do texto. Geralmente, essas palavras-chave são inseridas no título, *tags*, corpo do texto e URL (Patil-Swati *et al.*, 2013). Gouveia (2021) sugere que

a percentagem ideal de palavras-chave seja de 4,0 % do texto principal.

Ter o *layout* responsivo é outra condição importante para o algoritmo. No fundo, é a capacidade do website se adaptar aos vários dispositivos, como o tablet, o computador *desktop* e o smartphone (Di Fátima, 2023). Um website responsivo melhora a experiência dos utilizadores e, consequentemente, impacta a classificação do motor de pesquisa (Ziakis *et al.*, 2019).

Já o *Sitemap XML*, ou mapa do website, permite o rastreamento de todas as suas páginas e as suas atualizações (Dean, 2023; Ziakis *et al.*, 2019). Para Patel (2021), o *Sitemap XML* é ainda mais importante em websites que são constantemente atualizados, diminuindo o tempo que o algoritmo leva para rastrear a mudança.

Os especialistas também analisam a importância dos títulos das páginas web para alcançar a boa classificação (Ziakis *et al.*, 2019). É a partir dos títulos que o *PageRank* percebe a prioridade dos conteúdos. Patel (2021) explica que o título principal (H1) é um convite à leitura, mas que muitos subtítulos confundem o algoritmo.

Outro fator fundamental é a meta-descrição. Quando o utilizador precisa escolher entre os vários resultados de pesquisa, o resumo do conteúdo ajuda na tarefa. Más descrições sugerem que um website é de baixa qualidade (Ziakis *et al.,* 2019). Patel (2021) recomenda que a descrição tenha no máximo 160 caracteres.

Para garantir a qualidade do conteúdo, a Google dá prioridade à relevância do website (Romero-Sánchez e Barrios-Hernández, 2023). Muito embora a qualidade seja um valor subjetivo, especialistas indicam quais são as suas características (Gouveia, 2021; Patel, 2021; Shenoy e Prabhu, 2016). O texto deve ter ao menos 1.500 palavras, com título, subtítulos, citações, listas, imagens, links internos e externos.

Por outro lado, o *PageRank* penaliza as chamadas técnicas de *Black Hat SEO*. Essas práticas procuram melhorar artificialmente a classificação de um website (Gudivada *et al.*, 2015). O objetivo é fazer crer ao algoritmo que um website tem a autoridade que não tem, constituindo uma prática ilegal (Agushinta *et al.*, 2023).

3. Dados e métodos

Este artigo apresenta uma análise quantitativa dos fatores de SEO que têm mais influência em uma estratégia bem conseguida de otimização de websites. O ponto de partida é a já conhecida expressão “o conteúdo é rei”, de Bill Gates, que indica que o conteúdo da página web é o fator mais importante da equação.

As investigações empíricas sobre SEO, com análise de grandes volumes de dados, ainda são recentes (Drivas *et al.*, 2020) e o impacto dos fatores precisa ser melhor mapeado (Romero-Sánchez e Barrios-Hernández, 2023). Assim, a estratégia metodológica deste estudo baseia-se no plano pago da Ubersuggest, ferramenta mundialmente reconhecida no universo SEO (Faustino, 2019; Stephen, 2020).

A seleção da amostra foi realizada a partir da técnica *snowball*. Inicialmente, o website *MoneyLab* – reconhecido como uma das grandes referências nacionais em literacia financeira (Moneylab, 2019) – foi escolhido. Usando a função de *websites similares* da Ubersuggest, foram selecionadas todas as páginas web da mesma categoria. Assim, a amostra conta com os 12 websites portugueses de literacia financeira sugeridos pela ferramenta (Tabela I).

Website	*URL*
MoneyLab	https://moneylab.pt/
Contas Poupança	https://contaspoupanca.pt/
Finanças com Ella	www.financascomella.com/
Ekonomista	www.e-konomista.pt
Doutor Finanças	www.doutorfinancas.pt/
TaoFinance	https://taofinance.pt/
Economia Finanças	https://economiafinancas.com/
Como Economizar	https://comoeconomizar.net/
Finanças dos 90	www.financasdos90.com/
Cat Poupança	www.catpoupanca.pt/
A Tio Patinhas	https://atiopatinhas.com/
Dama de Ouros	www.damadeouros.com/

Tabela I. Websites portugueses em análise (n = 12)

A Ubersuggest permitiu a extração automatizada de um conjunto vasto de dados, como, por exemplo: autoridade do domínio, meta-descrições, *backlinks*, tráfego orgânico, velocidade do website, palavras-chave, pontuação *on-page*, entre outros. Porém, é importante reconhecer que fatores não analisados, como a experiência do utilizador (UX) ou as diversas métricas das redes sociais, também influenciam o desempenho dos websites nos motores de pesquisa. Os dados correspondem ao intervalo de um ano.

Cada website foi analisado conforme os fatores de SEO apontados na revisão de literatura como os mais importantes (Dean, 2023; Gouveia,

2021; Faustino, 2019). Além do exame holístico das páginas web, inclusive do seu código de programação, também foi mapeada a publicação com mais tráfego orgânico. É justamente esse artigo que permitiu encontrar os fatores relacionados à qualidade do conteúdo: tamanho do texto, densidade das palavras-chave, subtítulos, unicidade do conteúdo, links internos e externos, etc. (Patel, 2021).

Os dados foram extraídos e analisados a partir de um computador *desktop* com os seguintes aspetos técnicos: HP, Intel Core i7, Windows 8.1 de 64 bits, disco rígido SATA 500 GB, placa gráfica HD 5500, memória SD RAM de 4 GB e ecrã HD BrightView com retroiluminação WLED e 15,6 polegadas. O *browser* utilizado foi o Chrome – navegador com a maior quota de mercado nacional e estrangeira (Statcounter, 2023).

Alguns fatores de SEO foram analisados numa perspetiva da totalidade do website – inclusive o seu código-fonte –, enquanto outros só consideram a página mais visitada. Nesse caso, as informações representam o momento da extração, numa lógica de métodos digitais (Rogers, 2013). Por fim, é importante destacar que diversas ferramentas realizam a mesma extração de dados da Ubersuggest e podem indicar resultados diferentes, sendo esta uma limitação e viés na leitura dos resultados.

4. Resultados

O ponto inicial para analisar a estrutura de um website são os níveis de navegação. Assim, considera-se o número máximo de cliques entre a página inicial e a página mais distante. Todos os websites da amostra apresentam resultados positivos neste fator SEO, com entre dois e três cliques. Quanto mais cliques forem necessários, mais trabalho o algoritmo da Google tem para indexar essas páginas web.

Já os *backlinks* são um ponto de partida para analisar a autoridade do domínio. Essa autoridade corresponde à relevância de um website quando comparado aos demais. No fundo, os *backlinks* são todas as citações feitas a uma determinada página na internet, como, por exemplo, nas plataformas de redes sociais ou outros websites (Tabela II).

Mais do que a simples quantidade de *backlinks*, é importante analisar também a qualidade dessas ligações (Patel, 2021). Os dados indicam que não existe uma relação direta entre websites com mais *backlinks* (ligações externas) e com mais *nofollow* (ligações externas consideradas de baixa qualidade e que não passam autoridade a outras páginas).

Website	*Backlinks (n)*	*Nofollow (%)*
Doutor Finanças	305321	0,3
Economia Finanças	271704	1,4
Ekonomista	61104	38,3
Contas Poupança	12430	5,5
MoneyLab	9221	0,5
TaoFinance	4604	6,2
Cat Poupança	1923	0,0
Como Economizar	1042	7,5
Finanças com Ella	328	2,7
Finanças dos 90	7	57,1
A Tio Patinhas	6	33,3
Dama de Ouros	0	0,0

Tabela II. Backlinks versus Nofollow (n = 12)

Os três websites com mais *backlinks* são *Doutor Finanças*, *Economia Finanças* e *Ekonomista*, enquanto a maior percentagem de *nofollow* surge em *Finanças dos 90* (57,1%). Se os *backlinks* enviam para o website, é importante analisar como ele responde ao chamado. A Figura 1 revela, numa escala de 0 a 100, as velocidades de carregamento da amostra em computador e em dispositivos móveis.

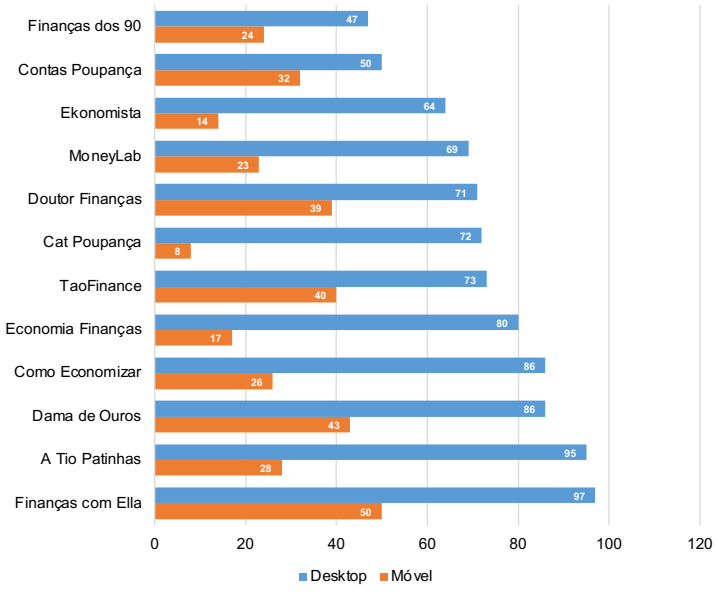

Figura 1. Velocidade de carregamento (0 a 100)

Os dados podem ser reunidos em três grupos, segundo a velocidade de carregamento no computador. A maioria dos websites tem mais de 50 pontos, considerado um desempenho razoável pela Ubersuggest. Os destaques positivos são,

respetivamente, *Finanças com Ella* (97) e *A Tio Patinhas* (95), com desempenhos de excelência. O único resultado francamente negativo foi encontrado para *Finanças dos 90*, abaixo dos 50 pontos.

Já o desempenho móvel apresenta resultados muito diferentes, segundo a velocidade de carregamento em smartphone e tablet. Neste tópico, a maioria dos websites tem uma classificação abaixo dos 50 pontos, considerados insuficientes. É possível notar a drástica diferença em *A Tio Patinhas* e *Como Economizar*, com resultados francamente inferiores.

Além da velocidade, é importante que o website tenha *layout* responsivo. Neste quesito, apenas o *Contas Poupança* não é *mobile friendly*. A amostra também se destaca na localização das páginas web. Todos os websites testados têm um *Sitemap XML*, sendo mais facilmente rastreados pelos motores de pesquisa (Patel, 2021).

Porém, não é só fundamental que as páginas web sejam mapeadas, mas também que os seus URLs estejam otimizados para pesquisa. Assim, para a leitura precisa dos resultados, é importante contabilizar as páginas bloqueadas por motores como o Google (Tabela III).

Website	*Páginas (n)*	*Bloqueadas (%)*
Como Economizar	404	63,0
Doutor Finanças	163	8,0
A Tio Patinhas	154	2,6
Contas Poupança	152	1,3
Dama de Ouros	84	2,4
Finanças com Ella	151	0,7
MoneyLab	150	0,0
Ekonomista	195	0,0
Economia Finanças	150	0,0
Finanças dos 90	150	0,0
TaoFinance	-	-
Cat Poupança	-	-

Tabela III. Páginas totais e bloqueadas pelo Google

Embora a maioria dos websites tenha percentagens nulas ou mesmo muito baixas, o *Como Economizar* destaca-se negativamente com 63,0% das páginas bloqueadas. É possível que o fenómeno seja criado por links partidos no código e que são penalizados pelos motores (Faustino, 2019). Problemas de segurança também poderiam justificar esse resultado.

Apesar de a idade do domínio ser um fator de SEO, é crucial examinar o seu histórico, refletido na autoridade do domínio (Dean, 2023). Páginas mais antigas tendem a ter mais prestígio no *ranking* dos motores de pesquisa. A Figura 2 revela, numa escala de 0 a 100, a autoridade do domínio da amostra. Quanto maior for o valor, melhor classificado estará o website.

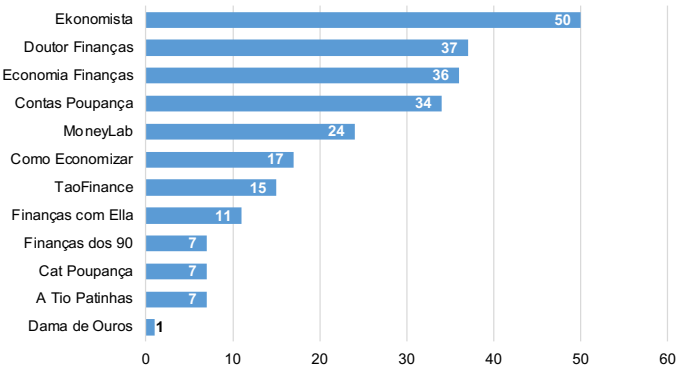

Figura 2. Autoridade de domínio (0 a 100)

Nenhum dos websites alcança um valor superior aos 50 pontos neste fator. O *Ekonomista* tem a maior autoridade, ainda assim com potencial de crescimento. É curioso notar que as outras páginas web mais bem classificadas aparecem com índices de influência semelhantes, com *Doutor Finanças* e *Economia Finanças*.

As redes sociais são um fator que pode alterar a equação, simplesmente porque representam hoje a origem mais frequente dos *backlinks*. Embora haja diferenças relevantes, os websites *Contas Poupança*, *Economia Finanças*, *Ekonomista* e *Doutor Finanças* têm o maior suporte no Facebook – a rede com a maior penetração em Portugal (OberCom, 2022). Estes também são os websites que mais publicaram os seus links na plataforma, com mais de 500 *posts* cada. Todos os websites restantes indicam volumes de interação baixos, ou mesmo nulos.

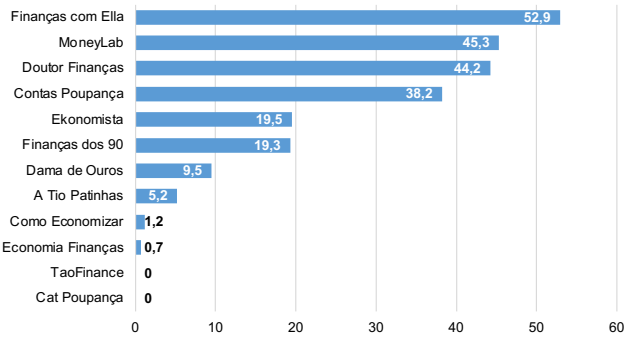

Figura 3. Títulos com erros de construção (%)

Os websites que se sobressaem nesse aspeto têm uma vantagem no *ranking* (Gudivada, 2015). É evidente que as publicações devem incluir o link do website para servirem como fonte de tráfego orgânico. Esses links estão vinculados ao título da página, que é um dos elementos mais relevantes em uma estratégia de otimização (Figura 3).

Os títulos dos websites podem apresentar anomalias, como serem muito curtos, longos ou ausentes, diminuindo o seu impacto. Neste ponto, o *Finanças com Ella* destaca-se negativamente (52,9 %), seguido por *MoneyLab* e *Doutor Finanças*. *TaoFinance* e *Cat Poupança* não têm dados, enquanto os demais websites apresentam menos de 20,0 % de erros. Além disso, páginas web sem meta-descrição perdem a oportunidade de atrair cliques (Figura 4).

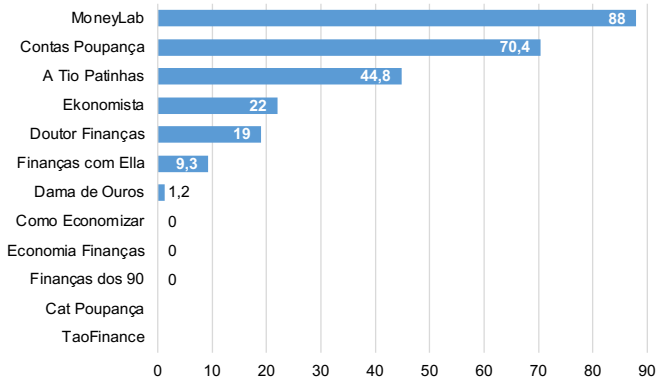

Figura 4. Páginas web sem meta-descrição (%)

A maioria da amostra tem páginas web sem meta-descrição. O caso mais grave é o *MoneyLab*, com 88,0%, seguido por *Contas Poupança* (70,4 %), *A Tio Patinhas* (44,8 %) e *Ekonomista* (22,1 %). Pelo lado positivo, destacam-se os websites *Finanças dos 90*, *Como Economizar* e *Economia Finanças*, com todas as suas páginas descritas.

Além dos fatores examinados na totalidade do website, há fatores que precisam ser analisados em uma página específica. Nesse caso, optou-se pela seleção da página mais visitada de cada website da amostra. Este é o caso dos links internos, presentes na metade das páginas com maior tráfego, dos *page headings* (41,7 %) e *alt text* (41,7 %).

Já para avaliar a qualidade do conteúdo, foram considerados seis parâmetros: número de palavras (> 1.500), estilização do texto, existência de conteúdo duplicado, atualização da página, presença de imagens e lista de itens. Assim, os websites poderiam ter de 0 a 6 pontos na classificação do *ranking* de qualidade (Tabela IV).

Website	*Parâmetros*	*Tráfego*
TaoFinance	6	1.081
Cat Poupança	6	229
Ekonomista	4	1.700.246
MoneyLab	3	1.246
Doutor Finanças	3	225.995
Economia Finanças	3	18.713
A Tio Patinhas	3	99
Dama de Ouros	3	22
Como Economizar	2	67
Finanças dos 90	2	11
Contas Poupança	1	54.073
Finanças com Ella	1	719

Tabela IV. Ranking de qualidade (parâmetros SEO versus tráfego orgânico)

A Tabela IV revela o cruzamento entre a soma dos aspetos relacionados na literatura especializada à qualidade do conteúdo (Gouveia, 2021; Patel, 2021; Ziakis *et al.*, 2019) e o tráfego orgânico dos websites. *TaoFinance* e *Cat Poupança* são os únicos que obedecem a todos os seis fatores, indicando que o seu conteúdo é considerado de maior qualidade. Por outro lado, *Contas Poupança* e *Finanças com Ella* atendem a apenas um fator positivo, sendo considerados de menor qualidade.

O *ranking* revela que não há uma relação direta entre a qualidade técnica do conteúdo e o tráfego orgânico. Websites muito distantes dos mais visitados da amostra têm o maior número de fatores relacionados à qualidade do conteúdo. Inclusive, o *Ekonomista* – website com mais tráfego orgânico – apresenta apenas quatro dos fatores de SEO.

5. Conclusões

Este artigo analisou a influência dos principais fatores de SEO em uma estratégia bem conseguida de otimização de media digitais. O ponto de partida foi a conhecida expressão “o conteúdo é rei”, dita por Bill Gates, em 1996, que indica claramente que o conteúdo da página web é o fator mais importante da equação.

A amostra é formada por 12 websites portugueses de literacia financeira, considerados de referência nacional. Os dados foram extraídos com a versão paga da Ubersuggest, e correspondem ao intervalo temporal de um ano.

Os *backlinks*, a autoridade do domínio e o suporte nas redes sociais são os fatores com o maior impacto no tráfego orgânico:

1. o website com mais tráfego (Ekonomista) tem a maior autoridade de domínio,
2. o segundo website mais visitado (Doutor Finanças) tem mais *backlinks*,
3. e o terceiro website com mais tráfego orgânico (Contas Poupança) tem o maior suporte nas redes sociais.

O fenómeno inverso também foi identificado em alguns websites da amostra, como no *Dama de Ouros*. Assim, menos *backlinks*, menor autoridade do domínio e pouco suporte nas redes sociais podem significar baixo tráfego orgânico.

Não existe uma relação direta entre a qualidade técnica do conteúdo e o tráfego orgânico. Esta conclusão vai de encontro a resultados de outros estudos, quanto à interligação entre o suporte nas redes sociais, a autoridade do domínio e os *backlinks* (Ziakis *et al.*, 2019).

Ao longo do tempo, os motores de pesquisa estabeleceram fatores para classificar a qualidade do conteúdo. Embora o próprio conceito de qualidade seja permeável a múltiplas análises, esses fatores impactam as etapas de rastreamento, indexação e criação do *ranking* dos algoritmos (Jain, 2013).

É importante destacar que diversas ferramentas realizam a mesma extração de dados da Ubersuggest e podem indicar resultados diferentes, sendo este um viés na leitura dos resultados. Porém, a principal limitação deste estudo é a impossibilidade técnica de avaliar os fatores de SEO relacionados ao tempo de visita no website e à taxa de rejeição dos utilizadores. Esses são dados privados, disponíveis apenas ao proprietário do domínio.

Surgir nos primeiros lugares dos motores de pesquisa não é uma tarefa fácil. Combina o funcionamento do algoritmo, o hábito do consumidor e o conteúdo do website. Não há uma fórmula mágica a ser seguida. A qualidade do conteúdo deveria refletir na autoridade, apesar de nem sempre ser isso o que acontece.

Embora, para atrair tráfego orgânico, o conteúdo em si não seja o fator mais importante de uma página web, apenas um bom conteúdo justifica o interesse dos utilizadores. Ou seja, como os motores de pesquisa são algorítmicos, eles podem não identificar os melhores conteúdos ou simplesmente refletir os vieses ideológicos dos seus construtores.

Referencias

Advanced Web Ranking (2021). Google organic CTR history: Fresh CTR averages pulled monthly from millions of keywords. // Advanced Web Ranking. https://www.advancedwebranking.com/ctrstudy/ (2023-08-21)

Agushinta, R. D.; Harmanto, S.; Suhendra, A.; Bastian, I.; Putrananda, V. A. (2023). The analysis of Indonesian regional websites with SEO methods. // Journal of Hunan University Natural Sciences. 50:5, 212-223. https://doi.org/10.55463/issn.1674-2974.50.5.20

Castells, M. (2012). Sociedade em rede. A era da informação: Economia, sociedade e cultura. Lisboa: Fundação Calouste Gulbenkian.

Chiarella, D.; Yarbrough, J.; Jackson, C. (2020). Using alt text to make science Twitter more accessible for people with visual impairments. // Nature Communications. 5803:2020, 1-3. https://doi.org/10.1038/s41467-020-19640-w

Dean, B. (2023). Google's 200 ranking factors: The complete list (2022). // Backlinko. https://backlinko.com/google-ranking-factors (2023-08-21)

Di Fátima, B. (2023). Depois do frenesi: Uma historiografia do jornalismo longform na internet. // Famecos. 30:1, e41773. https://doi.org/10.15448/1980-3729.2023.1.41773

Drivas, I. C.; Sakas, D. P.; Giannakopoulos, G. A.; Kyriaki-Manessi, D. (2020). Big data analytics for search engine optimization. // Big Data and Cognitive Computing. 4:2, 1-5. https://doi.org/10.3390/bdcc4020005

Duong, V. (2019). SEO management: methods and techniques to achieve success. Londres: Wiley.

Escandell-Poveda, R.; Papí-Gálvez, N.; Iglesias-García, M. (2023). Técnicas digitales para el estudio de las competencias y perfiles profesionales: el caso de la oferta laboral de SEO. // Scire. 29:1, 31–42. https://doi.org/10.54886/scire.v29i1.4877

Faustino, P. (2019). Marketing digital na prática. Lisboa: Marcador.

Feldmann, A.; Gasser, O.; Lichtblau, F.; Pujol, E.; Poese, I.; Dietzel, C.; Wagner, D.; Wichtlhuber, M.; Tapiador, J.; Vallina-Rodriguez, N.; Hohlfeld, O.; Smaragdakis, G. (2021). Implications of the Covid-19 pandemic on the Internet traffic. // Proceedings of the 15th ITG-Symposium, Online, 9 Abr., 2021. Broadband Coverage in Germany, 1-5. https://ieeexplore.ieee.org/document/9399711/authors#autho rs

Finch, S. (2019). Why Google is your most important learning tool. // People Management. https://www.peoplemanagement.co.uk/article/1744278/google-most-important-learning-tool (2023-08-21)

Gouveia, M. (2021). 20 Dicas de SEO para otimizares o teu website WordPress. // MarcoGouveia.PT. www.marcogouveia.pt/wordpress-seo/ (2023-08-21)

Gudivada, V. N.; Rao, D.; Paris, J. (2015). Understanding search-engine optimization. // Computer. 48:10, 43-52. https://doi.org/10.1109/MC.2015.297

Hootsuite. (2021). Digital 2021: Global overview report. // Hootsuite. https://datareportal.com/reports/digital-2021-global-overview-report (2023-08-21)

Ippolita (2013). The dark side of Google. // Institute of Network Cultures. https://networkcultures.org/blog/publication/no-13-the-dark-side-of-google-ippolita/

Jain, A. (2013). The role and importance of search engine and search engine optimization. // International Journal of Emerging Trends e Technology in Computer Science. 2:3, 99-102.

Lopezosa, C.; Codina, L.; Gonzalo-Penela, C. (2019). SEO off page y construcción de enlaces: estrategias generales y transmisión de autoridad en cibermedios. // El Profesional

de La Información. 28:1, 1-13. https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/66177

Lopezosa, C.; Guallar, J.; Santos-Hermosa, G. (2022). Google Discover: entre la recuperación de información y la curación algorítmica. // Scire. 28:2, 13-22. https://doi.org/10.54886/scire.v28i2.4796

MoneyLab. (2019). Bárbara Barroso eleita a Nº1 das Finanças em Portugal. // MoneyLab. https://moneylab.pt/2019/03/18/barbara-barroso-eleita-a-no1-das-financas-em-portugal

OberCom (2022). Digital News Report Portugal 2022 (Junho). // OberCom. https://obercom.pt/digital-news-report-2022-portugal/ (2023-08-21)

Ortega, M. C. (2020). Herramientas del marketing digital que permiten desarrollar presencia online, analizar la web, conocer a la audiencia y mejorar los resultados de búsqueda. // Perspectivas. 23:45, 33-60.

Palanisamy, R.; Liu, Y. User' search satisfaction in search engine optimization: An empirical analysis. // Journal of Services Research. 18:2, 83-120. http://dx.doi.org/10.1007/978-3-030-24643-3_124

Patel, N. (2021). What is SEO? Your complete step-by-step guide. // NeilPatel.Com. https://neilpatel.com/what-is-seo/ (2023-08-21)

Patil-Swati, P.; Pawar, B. V.; Patil-Ajay, S. (2013). Search engine optimization: A study. // Research Journal of Computer and Information Technology Sciences. 1:1, 10-13.

Pedrosa, L.; Morais, O. J. (2021). Visibilidad web en buscadores: factores algorítmicos de SEO on page (FAOP) como técnica e práctica periodística. // Estudios sobre el Mensaje Periodístico. 27:2, 579-591. https://doi.org/10.5209/esmp.71291

Rogers, R. (2013). Digital methods. Cambridge: The MIT Press.

Romero-Sánchez, D. F.; Barrios-Hernández, D. (2023). Adopción del comercio electrónico en el sector hortofrutícola: Un análisis en tiempos de pandemia. // Innovar. 33:87, 59-72. https://doi.org/10.15446/innovar.v33n87.105505

Sanchez-Cuadrado, S.; Morato, J. (2023). Análisis de respuestas enriquecidas en Google. // Scire: representación y organización del conocimiento. 29: 1, 13–23. https://doi.org/10.54886/scire.v29i1.4908

Sebring, S. S. (2019). Betting on SEO: The race to the top (of a Google search) isn't always as straight-forward as it seems. // CU Management. 42:5, 14-17.

Seymour, T.; Frantsvog, D.; Kumar, S. (2011). History of search engines. // International Journal of Management e Information Systems IJMIS). 15:4, 47-58. https://doi.org/10.19030/ijmis.v15i4.5799

Shenoy, A.; Prabhu, A. (2016). Introducing SEO: Your quick-start guide to effective SEO practices. Mumbai: Apress.

Statcounter. (2023). Top desktop, tablet e console browsers per country. // Statcounter. http://gs.statco unter.com/ (2023-08-21)

Stephen, G. (2020). Web analytics for the domain of Anna Centenary Library, Tamil Nadu: a study of using Ubersuggest tool. // Library Philosophy and Practice. https://digital commons.unl.edu/libphilprac/3671 (2023-08-21)

Yussuf, T. (2020). Search engine success. // The Home-Based Entrepreneur's Magazine, 30-31.

Ziakis, C.; Vlachopoulou, M.; Kyrkoudis, T.; Karagkiozidou, M. (2019). Important factors for improving Google search rank. // Future Internet. 11:2, 1-32. https://doi.org/10.3390/fi110200

Enviado: 2023-08-25. Segunda versión: 2024-06-06.
Aceptado: 2024-06-12.

Índice de autores

Author index

Índice de materias en español

Subject index in Spanish

Índice de materias en inglés

Subject index in English