



**Facultad de Comunicación y Documentación**

Grado en Gestión de Información y Contenidos Digitales

# **Transformación de metadatos OAI -PMH a conjuntos de datos semánticos RDF en revistas científicas y repositorios institucionales**

Autor: **Pedro Otálora Giménez**

Tutores:  
Juan-Antonio Pastor-Sánchez  
Tomás Saorín Pérez

Julio 2025



## Índice

Resumen.....	5
Abstract.....	5
1. Introducción.....	6
1.1 Estado de la cuestión.....	7
1.2 Justificación.....	9
1.3 Objetivos.....	10
2. Metodología.....	11
3. Resultados.....	12
3.1. Evaluación de las fuentes de datos.....	12
3.1.1. Esquemas de metadatos disponibles.....	13
3.1.2. Calidad de los metadatos.....	14
3.1.3. Acceso a los metadatos.....	15
3.1.4. Resumen de resultados del análisis de fuentes.....	15
3.2. Perfil de aplicación RDF.....	16
3.3. Proceso de transformación automatizado.....	18
3.3.1. Extracción de datos.....	21
3.3.2. Limpieza y normalización.....	22
3.3.3. Enriquecimiento semántico.....	25
3.3.4. Clasificación temática.....	26
3.3.5. Generación de ficheros RDF.....	30
3.4. Interfaz web de búsqueda y consulta.....	32
3.4.1. Arquitectura de la aplicación.....	32
3.4.2. Desarrollo de la aplicación.....	35
4. Conclusiones.....	44
5. Conclusions.....	47
6. Referencias bibliográficas.....	50
7. Anexos.....	54



## Índice de figuras

Figura 1. Modelo de representación para las revistas científicas del repositorio.....	16
Figura 2. Representación de vocabularios en el perfil de aplicación RDF.....	17
Figura 3. Diagrama de las etapas del proceso de transformación de datos.....	18
Figura 4. Entorno de ejecución en Google Colab y árbol de directorios de trabajo.....	19
Figura 5. Diagrama de vinculación con tesauros.....	26
Figura 6. Proceso de clasificación temática.....	27
Figura 7. Nube de palabras con la frecuencia de términos en el corpus artículos.....	29
Figura 8. Distribución temática de las revistas en el tesauro de la UNESCO.....	30
Figura 9. Diagrama del modelo RDF aplicado al conjunto de datos unificado.....	31
Figura 10. Diagrama de la arquitectura de la aplicación web.....	33
Figura 11. Interfaz web del servidor Apache Jena Fuseki.....	34
Figura 12. Diseño modular de componentes de la aplicación.....	35
Figura 13. Página de inicio de la aplicación web. Sección de estadísticas.....	37
Figura 14. Página de inicio de la aplicación web. Sección de rankings.....	38
Figura 15. Página de ficha de artículo. Sección de detalles.....	39
Figura 16. Página de ficha de artículo. Sección “artículos de la misma categoría”.....	40
Figura 17. Página de ficha de número.....	41
Figura 18. Página de listado de revistas.....	42
Figura 19. Detalle de la página de búsqueda de artículos y filtros.....	43
Figura 20. Detalle de la ficha de autor con tabla de co-autorías.....	43



## Índice de tablas

Tabla 1. Formatos de metadatos usados en Revistas UM y Digitum.....	13
Tabla 2. Detalle de características de metadatos en Revistas UM y Digitum.....	15
Tabla 3. Resumen de la ejecución del conjunto de scripts.....	20
Tabla 4. Archivos de revisión de autores.....	24
Tabla 5. Archivos de revisión de palabras clave generados en la normalización.....	24
Tabla 6. Ejemplo del texto de entrada en un registro y su salida tras el PLN.....	28
Tabla 7. Resultado de la técnica PLN en el corpus de un microtesauro.....	28
Tabla 8. Resultados de la creación del conjunto de datos.....	31
Tabla 9. Resumen de los módulos de la aplicación y ficheros de configuración.....	36

## Índice de anexos

Anexo I. Fragmento de Código del script de extracción de OAI-PMH a XML.....	54
Anexo II. Fragmento del Perfil de aplicación en formato TTL.....	55
Anexo III. Ejemplo de informes del proceso de transformación.....	56
Anexo IV. Archivos CSV intermedios creados en el proceso de limpieza.....	57
Anexo V. Fragmento de Dataset final en formato TTL.....	59
Anexo VI. Fragmento de Código PHP de la aplicación.....	60
Anexo VII. Glosario de términos y acrónimos.....	62

**Enlace al prototipo de aplicación web:** <http://gicd.inf.um.es/datarevistas/>

**Repositorio en GitHub:** <https://github.com/pedro-otalora/OAI2RDF>



Este trabajo está sujeto a una licencia de uso Creative Commons Atribución 4.0 Internacional (CC BY 4.0). Se permite cualquier explotación de la obra, así como la creación de obras derivadas, la distribución de las cuales también está permitida sin ninguna restricción salvo el reconocimiento de la autoría. Más información en: <https://creativecommons.org/licenses/by/4.0/>



## Resumen

Este proyecto aborda la transformación de metadatos OAI-PMH en conjuntos de datos semánticos RDF, en particular datos bibliográficos de revistas científicas y repositorios institucionales. Se desarrolló un proceso de transformación automatizado para transformar repositorios completos sin intervención del usuario. En el proceso se implementaron operaciones de clasificación temática basadas en técnicas de Procesamiento de Lenguaje Natural (PLN). Como caso de estudio, se ha aplicado el proceso al repositorio de revistas científicas de la Universidad de Murcia. El trabajo incluye el análisis de las fuentes de datos, la creación de un perfil de aplicación RDF, el diseño y ejecución del proceso de transformación de metadatos OAI-PMH a RDF, y la creación de una interfaz web para la explotación del conjunto de datos. Los resultados demuestran la viabilidad y eficacia del proceso, lo que facilita su integración en la Web Semántica.

**Palabras clave:** Repositorios institucionales. Revistas científicas. OAI-PMH. RDF. Web semántica. Interoperabilidad. Datos enlazados.

## Abstract

This project addresses the transformation of OAI-PMH metadata into semantic RDF datasets, specifically focusing on bibliographic data from scientific journals and institutional repositories. An automated transformation process was developed to convert entire repositories without requiring manual intervention. The process incorporates thematic classification operations based on Natural Language Processing (NLP) techniques. As a case study, the methodology was applied to the scientific journal repository of the University of Murcia. The work encompasses data source analysis, creation of an RDF application profile, design and execution of the OAI-PMH-to-RDF metadata transformation workflow, and development of a web interface for dataset exploration. The results demonstrate the feasibility and effectiveness of the process, facilitating its integration into the Semantic Web.

**Keywords:** Institutional repositories. Scientific journals. OAI-PMH. RDF. Semantic Web. Interoperability. Linked Data.



## 1. Introducción

En el escenario digital actual, el conocimiento científico se encuentra fragmentado en diferentes sistemas que funcionan de forma aislada; esto dificulta la relación entre la información existente y por tanto su descubrimiento. Esto representa un obstáculo para el aprovechamiento del conocimiento disponible, limitando la capacidad para establecer relaciones entre recursos (Berners-Lee et al., 2011). En esta misma línea, Cals et al. (2018) afirman que la Web Semántica da respuesta a este desafío con su tecnología. Su aplicación hace que la información pueda ser entendida y procesada por máquinas y por personas.

Por otra parte, los repositorios recopilan, difunden y preservan la producción intelectual, mejorando su accesibilidad y visibilidad (Abadal et al., 2013). Más allá de ser simples almacenes digitales, Ferreras-Fernández y Merlo-Vega (2015) los describen como instrumentos estratégicos que cumplen con las políticas de acceso abierto y garantizan la preservación a largo plazo del patrimonio académico. De hecho, Morcillo López (2016) los cataloga como aliados de las universidades para lograr el acceso abierto mediante el auto-archivo, también denominado “ruta verde” de publicación.

Entre los contenidos de los repositorios institucionales destacan las revistas científicas de acceso abierto, instrumentos fundamentales en la difusión del conocimiento y desarrollo disciplinar (Osca-Lluch et al., 2008). En el ámbito universitario, estas publicaciones dan visibilidad a la producción científica institucional (Deroy Domínguez, 2022).

En el contexto español, tras el auge de los de repositorios a principios de siglo, autores como Melero (2008) ya concluyeron que la herramienta más utilizada para su gestión era –y lo sigue siendo en la actualidad– DSpace, una plataforma de código abierto creada específicamente para la gestión de repositorios. De forma paralela, muchas universidades ofrecen las revistas académicas en portales web publicados con OJS (Open Journal System). Aunque la función principal de OJS es la gestión del flujo editorial, también incorpora opciones de publicación propias de los repositorios.



El conjunto de repositorios, sin una conexión entre ellos, suponen fuentes de datos aislados. Para dar solución a este aislamiento, los repositorios adoptan mayoritariamente el protocolo OAI-PMH –desarrollado por Open Archives Initiative– como estándar de interoperabilidad. El protocolo OAI-PMH hace posible que plataformas denominadas recolectores y agregadores agrupen los contenidos de los repositorios y actúen como fuentes de datos agregados. En concreto, en el caso de Hispana y Europeana, el alto grado de calidad alcanzado en sus colecciones radica en el uso de Linked Open Data y la ontología Europeana Data Model (Xavier y Hernández, 2020).

Pero el modelo de datos enlazados, más allá de organizar colecciones, es una herramienta que permite articular los datos en grafos de conocimiento. Por ello, autores como Piedra et al. (2015) opinan que la transformación de metadatos hacia formatos semánticamente enriquecidos resulta crucial. Es decir, para superar las limitaciones actuales es necesario evolucionar hacia un modelo semántico que supere los sistemas tradicionales de recuperación de información (Martínez-Ávila, 2018).

## 1.1 Estado de la cuestión

La gestión de datos de investigación en repositorios universitarios españoles presenta importantes desafíos. Un estudio reciente de Monteagudo-Haro y Prieto-Gutiérrez (2024) revela que apenas el 52% de estas instituciones proporcionan acceso a datos científicos, con marcadas diferencias según su financiación: el 73% en universidades públicas frente al 12% en privadas. Esta disparidad evidencia una brecha crítica en las políticas de acceso abierto, requiriendo soluciones que optimicen la publicación y conexión de conjuntos de datos académicos. Atendiendo a estas cifras, los beneficios afectarían especialmente a las instituciones públicas.

Con carácter general, proyectos como la iniciativa Hércules<sup>1</sup> proponen soluciones para crear un sistema integral de gestión de la investigación con capacidades semánticas (Universidad de Murcia, s.f.). Hércules se basa en datos abiertos y busca crear, mediante ontologías, una estructura semántica para la información procedente de la investigación que permita la interoperabilidad entre distintas instituciones. Siguiendo las

---

<sup>1</sup> Más información sobre el proyecto Hércules en <https://www.um.es/web/hercules/inicio>.



buenas prácticas para los datos enlazados, el proyecto Hércules apuesta por la reutilización de vocabularios existentes en el ámbito académico como BIBO, VIVO, AlISO, etc.

El portal de datos abiertos Aragón Open Data<sup>2</sup> es otra muestra de la mejora de calidad al utilizar Linked Open Data. En Aragopedia.org se pueden consultar y relacionar datos de distintos tipos de fuentes a través de grafos de conocimiento que proporcionan las tecnologías semánticas (Gobierno de Aragón, 2023). Aunque este portal recoge datos del territorio aragonés de ámbitos distintos al académico, el acuerdo de colaboración con la Universidad de Zaragoza (Unizar) pone a disposición los conjuntos de datos de su repositorio<sup>3</sup> para ser incorporados de forma automática. La integración se realiza a través del software de código abierto CKAN, un gestor de contenidos para datos abiertos. Mediante una adaptación *ad hoc*, los metadatos OAI-PMH son incorporados de forma automática al portal Aragón Open Data (Alcober Fuertes, s.f.).

Un caso más específico es la revista argentina Ciencia y Técnica Administrativa - Cyta<sup>4</sup>, que utiliza el software OJS para gestionar y publicar los contenidos. Los editores de la revista han desarrollado un proceso de transformación de datos en formato OAI-PMH a RDF (Resource Description Framework). Utilizando una metodología de extracción basada en MySQL, realiza la conversión de datos usando scripts PHP, ofreciendo el conjunto de datos para su descarga en formato TTL (formato de datos abiertos que destaca por su sintaxis simple y legible). Este conjunto de datos también se encuentra disponible a través de un punto de acceso SPARQL donde se pueden realizar consultas estructuradas. Aunque en la actualidad no ofrece ninguna aplicación específica para explorar el conjunto de datos RDF generado, el punto de acceso ofrece la posibilidad de realizar búsquedas federadas enlazando el grafo de conocimiento generado con otras fuentes externas, ampliando enormemente la interoperabilidad original ofrecida por OAI-PMH.

Dejando atrás trabajos para entornos específicos, el proyecto LOD-GF de la Universidad de Cuenca<sup>5</sup> ofrece una solución más alineada con los objetivos de este

---

<sup>2</sup> Acceso al portal Aragón Open Data: <https://opendata.aragon.es/>

<sup>3</sup> Repositorio de la Universidad de Zaragoza: (<https://zaguan.unizar.es>)

<sup>4</sup> Más información acerca del repositorio OAI-PMH de la revista Cyta: <https://www.cyta.com.ar/oai/>

<sup>5</sup> Repositorio del proyecto LOD-GF en GitHub: <https://ucuenca.github.io/lodplatform/>



trabajo. Utilizando la herramienta de Business Intelligence Pentaho, el proyecto ha creado un framework completo de transformación de datos OAI-PMH a formato RDF que incluye procesos de modelado, limpieza de datos, publicación y explotación del conjunto de datos resultante. El uso de plugins nativos de Pentaho Data Integration así como otros generados específicamente, permite realizar el proceso a través de una interfaz gráfica utilizando mecánicas amigables para el usuario. El sistema de desambiguación de autores implementado a través de Silk Workbench<sup>6</sup> es uno de los puntos fuertes del proyecto, ya que es capaz de detectar duplicidades utilizando los títulos de sus obras para realizar un proceso de vinculación preciso.

## 1.2 Justificación

La aplicación de técnicas semánticas en repositorios institucionales genera impactos medibles. Peset et al. (2017) documentan aumentos del 30-40% en la visibilidad de contenidos académicos tras implementar estas tecnologías. Estos sistemas permiten conectar investigaciones relacionadas automáticamente, creando redes de conocimiento interdisciplinares.

En este sentido, la transformación de datos bibliográficos a RDF es una estrategia clave para mejorar la difusión científica universitaria. Este proceso facilita la integración de fuentes heterogéneas bajo un estándar común de interoperabilidad. No obstante, su implementación práctica enfrenta obstáculos significativos.

El principal desafío radica en la importante carga de trabajo manual requerida. Automatizar estas tareas, particularmente en repositorios basados en OAI-PMH, podría contribuir en la integración de datos académicos en modelos semánticos. El presente trabajo aborda este reto mediante un sistema específico para metadatos de revistas científicas, incorporando dos avances innovadores: clasificación temática automática usando el tesoro UNESCO con técnicas de PLN y una interfaz web que potencia la exploración conceptual mediante relaciones semánticas.

Para validar este proceso, se ha seleccionado el repositorio de la Universidad de Murcia como caso de estudio<sup>7</sup>. Este repositorio alberga 67 revistas, de las cuales 43 se

---

<sup>6</sup> Sitio web de Silk Workbench: <http://silkframework.org/>

<sup>7</sup> Portal web de Revistas UM: <https://revistas.um.es>



encuentran en activo (Revistas UM, s.f.), constituyendo un corpus representativo. Su extensión permite evaluar tanto la eficacia del proceso de transformación como la precisión en la categorización temática automatizada.

### 1.3 Objetivos

**Objetivo General:** Desarrollar un proceso de transformación automatizado para la creación de conjuntos de datos RDF a partir de repositorios que utilizan OAI-PMH y validar el proceso utilizando como caso de estudio el repositorio de revistas científicas de la Universidad de Murcia.

#### Objetivos Específicos:

1. Analizar la estructura y los metadatos de la fuente OAI-PMH, en este caso, el repositorio de revistas científicas de la Universidad de Murcia.
2. Desarrollar un perfil de aplicación RDF que represente el dominio de las revistas científicas, reutilizando vocabularios estándar como BIBO o Dublin Core, que permitan la interoperabilidad del conjunto de datos.
3. Diseñar e implementar un proceso de transformación para extraer, limpiar y transformar los metadatos del repositorio OAI-PMH en un conjunto de datos RDF, aplicando procesos de normalización y enriquecimiento con la mayor autonomía posible y que permita su escalabilidad en distintos escenarios.
4. Incorporar en el proceso de transformación técnicas de PLN para conseguir una clasificación temática automatizada.
5. Desarrollar una interfaz web que permita la navegación y el descubrimiento de los artículos científicos mediante búsquedas facetadas, a través de un punto de acceso SPARQL.



## 2. Metodología

La presente investigación adopta un enfoque empírico para la transformación de metadatos bibliográficos en conjuntos de datos enlazados. Se desarrolló una metodología que integra aspectos cuantitativos y cualitativos, cuyo objeto de estudio es el ciclo completo de transformación de metadatos. Para validar la propuesta, se eligió como caso de estudio el repositorio Revistas UM. Esta elección permitió evaluar tanto la viabilidad técnica del flujo de transformación como la calidad de los metadatos hallados en el repositorio Revistas UM.

La metodología consta de cuatro pasos: la identificación y análisis de fuentes de datos, la creación del perfil de aplicación RDF, el proceso de transformación de metadatos OAI-PMH a un conjunto de datos semánticos, y la interfaz web para la explotación del conjunto de datos.

**Paso 1. Identificación de fuentes de datos.** Se localizaron y analizaron los repositorios OAI-PMH relevantes para el caso de estudio, seleccionando las fuentes primarias para el ciclo de transformación.

**Paso 2. Creación del perfil de aplicación RDF.** Se diseñó un perfil de aplicación RDF para representar las entidades y relaciones del dominio de revistas científicas, utilizando vocabularios estándar y validando la estructura en una plataforma especializada.

**Paso 3. Transformación de metadatos OAI-PMH a RDF.** Se desarrolló un proceso de transformación automatizado para extraer, limpiar, enriquecer y clasificar los metadatos de un repositorio OAI-PMH, generando a su salida un conjunto de datos semánticos RDF, siguiendo el perfil de aplicación definido.

**Paso 4. Desarrollo de la interfaz web de exploración y búsqueda.** Se desarrolló una aplicación web para la consulta y visualización de los datos. Para su implementación se siguió una estructura modular basada en la arquitectura de navegación, utilizando el modelo cliente-servidor apoyado en un punto de acceso SPARQL para el acceso al conjunto de datos.



### 3. Resultados

En esta sección se presentan los resultados obtenidos durante las diferentes fases del trabajo. En ellos se incluye el análisis inicial de fuentes y el estudio del perfil de aplicación RDF. Los resultados del proceso de transformación automatizado son los procedentes de su aplicación sobre el repositorio Revistas UM.

#### 3.1. Evaluación de las fuentes de datos

En el proceso de identificación de las fuentes OAI-PMH de Revistas UM se llevó a cabo una exploración de los agregadores Recolecta, Hispana y Europeana. Tras la exploración se determinó que los distintos hallazgos de Revistas UM en los agregadores conformaban fuentes de información secundaria, por lo que se optó por analizar las fuentes primarias referenciadas en los agregadores: el repositorio Digitum (repositorio de la Universidad de Murcia) y el propio portal web de Editum (sello editorial de la Universidad de Murcia).

Tras una primera exploración de las dos fuentes primarias se comprobó que cada repositorio utilizaba una plataforma de publicación distinta, ofreciendo ambas una información similar, pero con diferentes características.

En el caso de Revistas UM, su portal web utiliza la plataforma OJS (Open Journal System), un software de gestión editorial de código abierto ampliamente usado en las publicaciones académicas. Se encontraron puntos de acceso independientes para cada revista científica, lo que facilitó posteriormente la extracción segmentada de datos.

Por otra parte, Digitum es gestionado con la herramienta DSpace, una plataforma de código abierto con características específicas para la gestión de repositorios. Esta fuente no contiene exclusivamente las revistas publicadas por Editum, sino que alberga otras colecciones relacionadas con la institución, como congresos, tesis, etc.

Para realizar la comparativa se realizó la extracción de muestras en formatos XML y CSV de registros correspondientes a los mismos recursos en ambos repositorios. El análisis comparativo entre Revistas UM (OJS) y Digitum (DSpace) permitió identificar diferencias sustanciales en formatos, calidad y estructura de metadatos ofrecidos por



ambos repositorios. Para medir los resultados se utilizaron tres indicadores: esquemas de metadatos disponibles, calidad de los datos y puntos de acceso.

### 3.1.1. Esquemas de metadatos disponibles

En el análisis de esquemas de metadatos se observó que Digitum ofrece una amplia variedad de formatos como uketd\_dc, mods o mets. Por su parte, Revistas UM presenta un conjunto más limitado: oai\_dc, oai\_marc, marcxml y rfc1807. En la Tabla 1 se presenta una comparativa con la cantidad de elementos disponibles para cada esquema de metadatos y el uso promedio que hace cada plataforma.

**Tabla 1.** Formatos de metadatos usados en Revistas UM y Digitum.

Formato de metadatos	Elementos disponibles en el formato	Elementos usados en Revistas UM	Elementos usados en Digitum
oai_marc	20	12	
marcxml	20	13	
rfc1807	12	11	
<b>oai_dc</b>	<b>15</b>	<b>9</b>	<b>11</b>
qdc	40		14
didl	21		15
mods	20		9
ore	13		10
mets	Variable		15
rdf	Variable		22
marc	20		19
xoai	Variable		27
dim	Variable		31
etdms	14		11

En el análisis detallado de los formatos disponibles en ambos repositorios se observó que comparten únicamente un esquema de metadatos común: oai\_dc, basado en Dublin Core. Esta diferencia en la tipología de esquemas ofrecidos por cada repositorio refleja una mayor orientación de DSpace para la preservación digital frente a OJS, que prioriza los formatos enfocados en la gestión de recursos bibliográficos.



### 3.1.2. Calidad de los metadatos

Por referencia directa al contexto de este trabajo, se examinó el formato de metadatos RDF hallado en DSpace. Tras visualizar su contenido, se observó que el contenido no aprovecha las capacidades semánticas de RDF/OWL. Se limita únicamente a ofrecer los metadatos Dublin Core sin añadir ningún contenido semántico propio de los datos enlazados.

El análisis del formato oai\_dc, coincidente en ambos repositorios, mostró un promedio de uso de elementos disponibles del 73% en Digitum frente al 60% de Revistas UM. Aunque Digitum presentó un mayor porcentaje de metadatos utilizados, un examen detallado permitió identificar dos aspectos relevantes para este trabajo:

- En primer lugar, el metadato dc:source no se encuentra presente en los registros de Digitum. Este metadato es necesario para establecer la jerarquía de los artículos con otras entidades en el proceso de transformación.
- En segundo lugar, Digitum carece del atributo de idioma en los metadatos. La ausencia de este elemento dificulta el filtrado por idioma del recurso, lo cual afecta negativamente la clasificación temática basada en técnicas de PLN.

En la Tabla 2 se muestra un fragmento del mismo recurso en cada repositorio. Se puede observar que Revistas UM incorpora el elemento “xml:lang” y el metadato <dc:source>, mientras que el registro en Digitum no contiene dicha información. También se puede observar que el metadato <dc:source> en Revistas UM se utiliza para indicar no solamente la revista y número a los que pertenece el artículo, sino que contiene múltiples repeticiones con datos del ISSN, ISSNE, DOI y otros. El hallazgo de este metadato como multivalorado, requirió de un tratamiento específico posterior para su normalización.



**Tabla 2.** Detalle de características de metadatos en Revistas UM y Digitum.

Metadatos en Revistas UM
<dc:title xml:lang="es-ES">Brucelosis bovina: estudio comparativo entre [...]</dc:title>
<dc:title xml:lang="en-US">Bovine brucellosis: a comparative study between [...]</dc:title>
<dc:subject xml:lang="es-ES">brucellosis</dc:subject>
<dc:source xml:lang="es-ES">Anales de Veterinaria de Murcia; Vol. 1 (1985); 5-16</dc:source>
<dc:source>1989-1784</dc:source>
<dc:source>0213-5434</dc:source>
Metadatos en Digitum
<dc:title>Brucelosis bovina: estudio comparativo entre ...</dc:title>
<dc:title>Bovine brucellosis: A comparative study between ...</dc:title>
<dc:subject>Brucellosis veterinaria</dc:subject>
<dc:source> (No presente en el registro)

### 3.1.3. Acceso a los metadatos

En Revistas UM se halló una URL específica para cada revista, lo que facilitó la extracción segmentada de los artículos. Por otra parte, en Digitum se observó un modelo basado en colecciones, las cuales no disponen de una jerarquía revista → volumen → artículo. Esto hace necesario conocer el código correspondiente de comunidad para cada número/volumen. Tras el análisis de las cabeceras XML de los registros, la etiqueta <setSpec> (etiqueta que contiene el valor de la comunidad) tampoco ofreció datos que permitieran identificar de manera sencilla la procedencia del registro.

Ambos repositorios presentaron la necesidad de utilizar técnicas de *web scraping* para conformar los puntos de acceso. Tras la configuración hallada en los portales, se concluyó que en el caso de Revistas UM era necesario extraer los puntos de acceso para cada revista (43 puntos de acceso), mientras que la fuente Digitum hubiese necesitado una extracción más compleja debido a la necesidad de extraer los puntos de acceso para cada volumen (más de 1.500 puntos de acceso distintos).

### 3.1.4. Resumen de resultados del análisis de fuentes

Como se ha observado, el análisis mostró que, mientras que Digitum ofrece una mayor variedad y cobertura de esquemas de metadatos, Revistas UM presenta una estructura

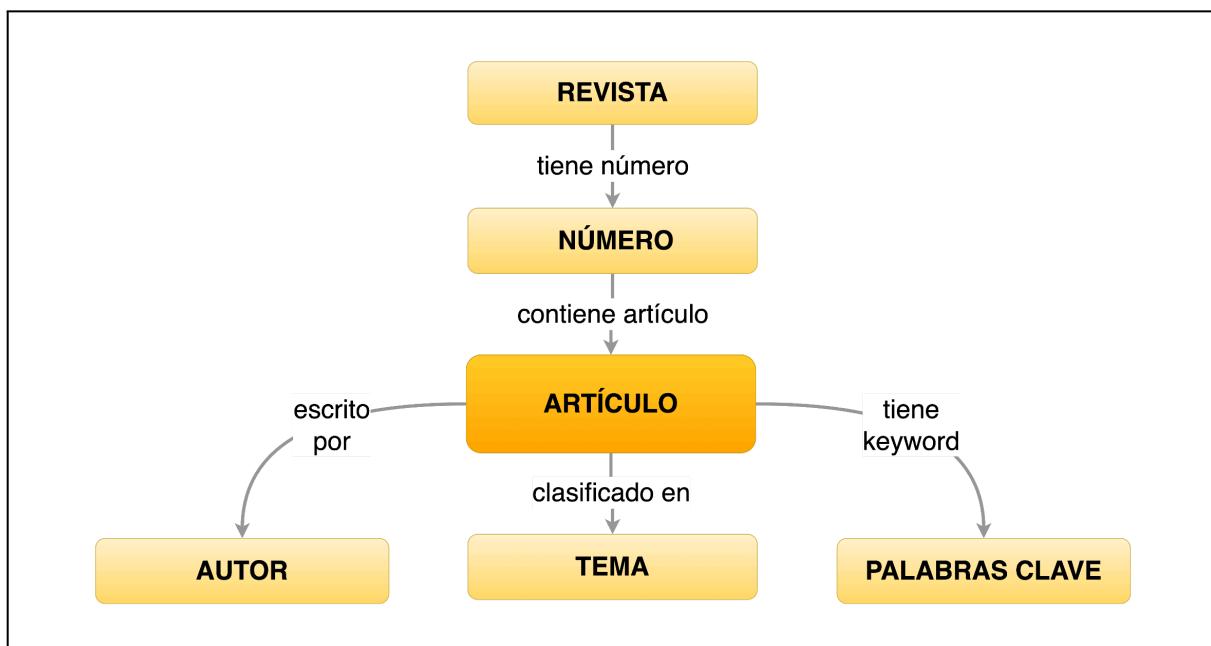


más limitada pero mejor adaptada a la publicación bibliográfica. En términos de calidad, aunque Digitum utiliza más elementos del formato oai\_dc, carece de metadatos clave como dc:source y elementos de etiquetado de idioma, información que sí se encuentra disponible en los registros de Revistas UM. Finalmente, el acceso a los metadatos resultó más sencillo y jerarquizado en Revistas UM, mientras que en Digitum la segmentación por comunidades y la falta de claridad en las etiquetas dificultaron la identificación de los registros.

### 3.2. Perfil de aplicación RDF

En el análisis de los registros OAI-PMH del repositorio Revistas UM, se identificaron los metadatos susceptibles de representar entidades (recursos identificables) y propiedades (atributos descriptivos). La estructura Revista → Número → Artículo refleja la organización propia de las publicaciones científicas, donde cada revista contiene números periódicos que a su vez agrupan artículos relacionados con autores, palabras clave y temas específicos. En el diagrama de la Figura 1 se puede observar la estructura jerárquica de las entidades identificadas.

Figura 1. Modelo de representación para las revistas científicas del repositorio.



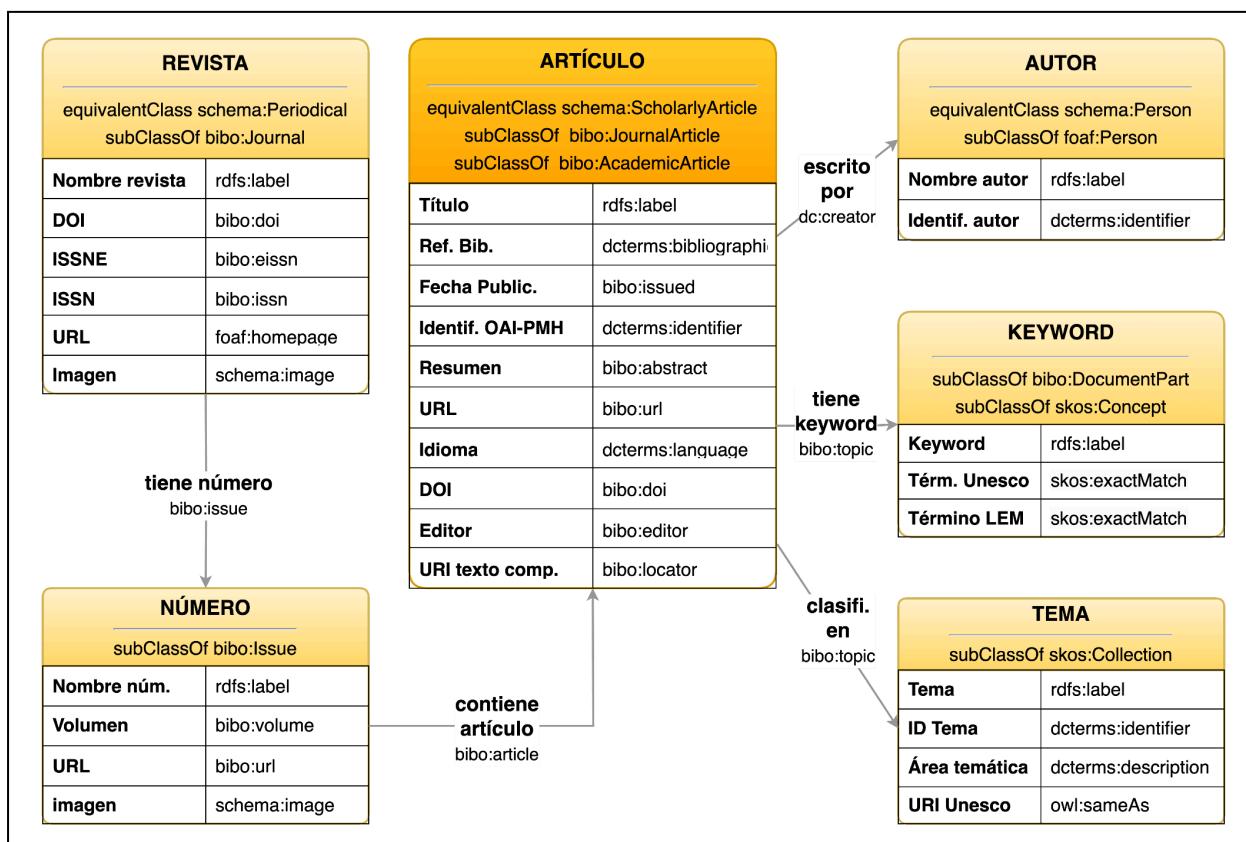
En la elaboración del perfil de aplicación se buscaron vocabularios estándar para describir recursos bibliográficos. Para la búsqueda se utilizó el portal LOV (Linked Open Vocabularies), donde se puede acceder a una amplia colección de vocabularios de



calidad. Se eligió BIBO (Bibliographic Ontology) como vocabulario principal ya que incluye términos específicos para describir revistas (bibo:Journal), números periódicos (bibo:Issue), artículos (bibo:Article) o propiedades para identificadores como bibo:doi y bibo:issn.

Adicionalmente, se incluyeron otros vocabularios como Skos, DCTerms, Foaf y Schema para describir aspectos específicos como términos de tesauro, personas o imágenes. Para mejorar la interoperabilidad se establecieron equivalencias entre vocabularios mediante las relaciones equivalentClass y equivalentProperty. Por ejemplo, la clase ontorevistas:Revista, que representa revistas científicas, se declaró equivalente a schema:Periodical mediante la propiedad owl:equivalentClass. Esto permite que las instancias de esta clase sean reconocidas como publicaciones periódicas tanto en este perfil de aplicación como en otros basados en Schema.org. La Figura 2 muestra el esquema general de entidades con las distintas propiedades de datos y relaciones, así como los vocabularios empleados para cada elemento.

Figura 2. Representación de vocabularios en el perfil de aplicación RDF.



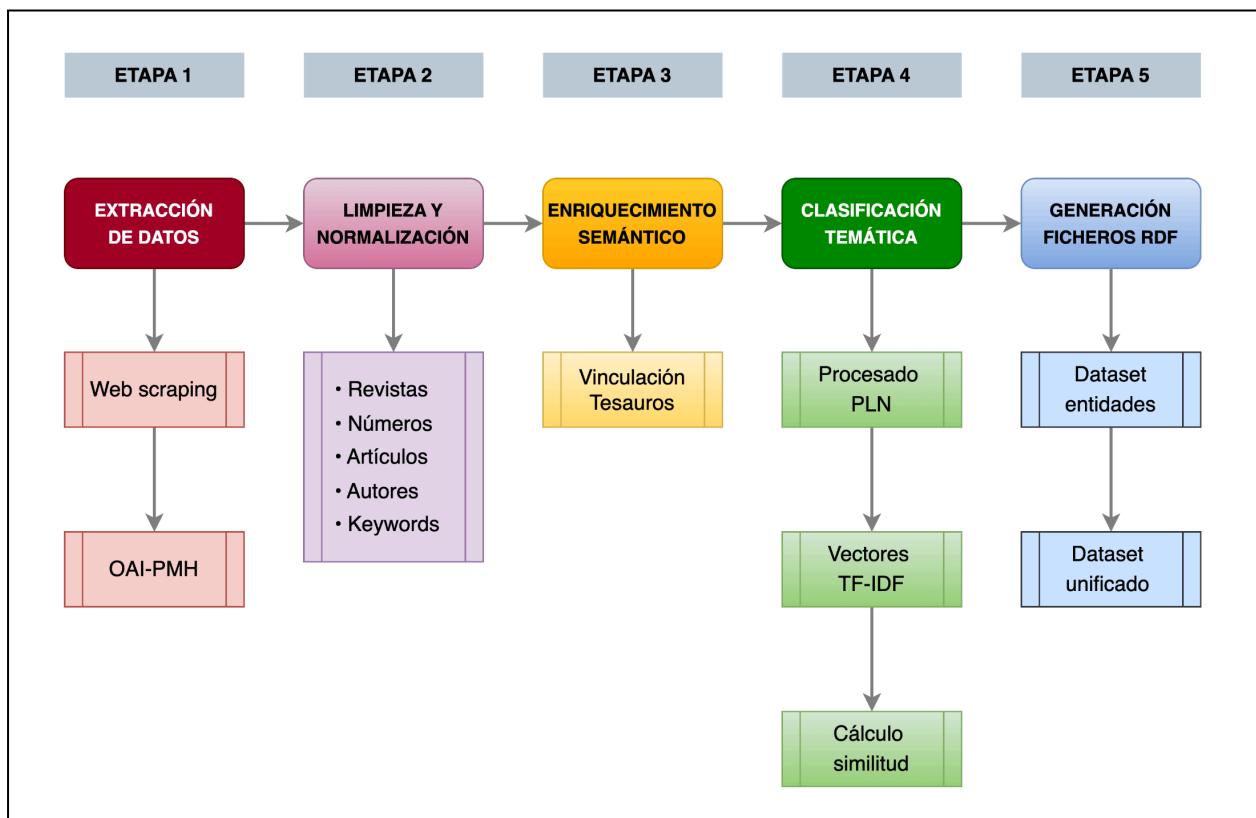


El esquema creado fue implementado en la herramienta Vocbech donde se pudo validar la estructura, relaciones y propiedades de las clases definidas. En el [Anexo II](#) se presenta un fragmento del perfil de aplicación en formato TTL. También se puede acceder al perfil completo en el repositorio [Github](#) del proyecto.

### 3.3. Proceso de transformación automatizado

Tras definir el perfil de aplicación RDF, se procedió a implementar el proceso de transformación automatizado, cuyos resultados se detallan a continuación. El proceso de transformación se dividió en cinco etapas que se ejecutan de manera secuencial, comenzando con la extracción de los datos del repositorio OAI-PMH hasta la obtención del conjunto de semánticos en formato TTL. En la Figura 3 se incluye un diagrama de las etapas y principales tareas que se ejecutan en cada una de ellas.

**Figura 3.** Diagrama de las etapas del proceso de transformación de datos OAI-PMH a RDF.



Para la implementación del proceso se utilizó el lenguaje de programación Python, elegido por su amplio ecosistema de librerías disponibles. Como entorno de programación se utilizó la plataforma Google Colab, donde se agruparon los distintos



módulos en un mismo proyecto. La Figura 4 contiene una captura de pantalla de Google Colab donde se puede observar el entorno de ejecución que proporciona la plataforma.

**Figura 4.** Entorno de ejecución en Google Colab y árbol de directorios de trabajo.

The screenshot shows a Google Colab interface with a notebook titled 'TFG\_PIPELINE\_V4.ipynb'. The left sidebar displays a file tree with directories like 'Archivos', 'GRAFICOS', 'LOGS', 'RDFOUT' containing files such as 'articulos.ttl', 'autores.ttl', 'dataset\_unificado.ttl', 'grupos\_tematicos.ttl', 'numeros.ttl', 'palabras\_clave.ttl', 'revistas.ttl', 'REVISION', 'TABLAS', 'TESAUROS', and 'XML'. The main area contains code for creating an RDF dataset from a OAI-PMH repository. The code includes importing libraries, mounting Google Drive, defining pipeline variables, and listing global variables.

```
CREACION DE UN DATASET RDF A PARTIR DEL REPOSITORIO OAI-PMH  
REVISTAS UM  
VERSION 4  
0. MONTAR DIRECTORIO DE ARCHIVOS  
0.1 Monta la carpeta de Drive y establece la variable PIPELINE con el directorio de trabajo  
[1] ## 0.1 Monta la carpeta de Drive y establece la variable PIPELINE con el directorio de trabajo  
# importar librerías  
from google.colab import drive  
  
# montar Google Drive en Colab  
drive.mount('/content/drive')  
  
# ruta de archivos del proyecto  
PIPELINE = "/content/drive/MyDrive/GICD/TFG/PIPELINE_V4/"  
TABLAS = PIPELINE + "TABLAS/"  
LOGS = PIPELINE + "LOGS/"  
XML = PIPELINE + "XML/"  
REVISION = PIPELINE + "REVISION/"  
GRAFICOS = PIPELINE + "GRAFICOS/"  
TESAUROS = PIPELINE + "TESAUROS/"  
RDFOUT = PIPELINE + "RDFOUT/"
```

Este proceso se diseñó para completar la transformación sin necesidad de intervención por parte del usuario. Para ello se realizó una configuración previa para ajustar los subprocessos al repositorio Revistas UM.

La definición de variables globales permitió la ejecución en bloque de todos los módulos que conforman el proceso, obteniéndose al finalizar el conjunto de datos RDF. Además, se incluyó un módulo encargado de preparar los directorios y descargar los ficheros de tesauros necesarios, dejando el entorno de ejecución listo para trabajar. En la Tabla 3 se presentan los módulos que componen el proceso de transformación, listados en orden de ejecución, junto con un resumen de los resultados obtenidos y el tiempo de ejecución empleado por cada uno de ellos.



Tabla 3. Resumen de la ejecución del conjunto de scripts.

Etapa	Resultado	Ejecución
<b>Extracción de datos</b>		
Extraer tabla de revistas ( <i>web scraping</i> )	43 revistas	0:00:33
Extraer artículos en XML (OAI-PMH)	24.650 registros	0:23:07
<b>Limpieza y normalización</b>		
Crear archivo artículos/volumenes.csv	1.596 volúmenes	> 0:00:01
Crear archivo números.csv	100% asignación artículos	> 0:00:01
Crear archivo artículos.csv	95% títulos en castellano	0:00:39
Crear archivo autores.csv	35.109 autores	0:02:35
Crear archivo autores_normalizados.csv	33.212 autores (-6%)	0:00:27
Crear archivo keywords.csv	39.892 palabras clave	0:02:43
<b>Enriquecimiento semántico</b>		
Reconciliación tesoro UNESCO	1.874 términos (5%)	0:00:07
Reconciliación LEM	2.987 términos (7%)	0:00:33
<b>Clasificación temática</b>		
Crear archivo corpus.csv de artículos	58% reducción de palabras	0:15:57
Crear archivo corpus.csv del tesoro	88 categorías	0:00:39
Crear archivo clusters.csv (con trazabilidad)	18.985 artículos clasificados (77%)	0:32:00
<b>Generación de ficheros RDF</b>		
Crear grupo_tema.ttl	533 tripletas	> 0:00:01
Crear revistas.ttl	321 tripletas	> 0:00:01
Crear números.ttl	31.034 tripletas	0:00:22
Crear autores.ttl	148.603 tripletas	0:00:29
Crear palabras_clave.ttl	164.031 tripletas	0:01:24
Crear artículos.ttl	257.195 tripletas	0:03:13
Crear dataset unificado.ttl	781.027 tripletas	0:01:17
		<b>Tiempo total</b>
		<b>1:26:08</b>

Durante la ejecución, cada módulo generó un informe con los hallazgos más relevantes del proceso: tiempo de ejecución, elementos procesados, resumen de resultados, etc. Estos informes se guardaron en ficheros de texto para su posterior revisión, a fin de controlar el resultado final y determinar si es necesaria alguna acción adicional en la normalización de los datos. En el [Anexo III](#) se muestran algunos ejemplos de las auditorías más relevantes generadas y almacenadas en la carpeta “logs”, ubicada dentro del directorio del proyecto.



Como se aprecia en los resultados, el proceso necesitó algo menos de 90 minutos para completar la transformación del repositorio Revistas UM. En el [Anexo IV](#) se muestran algunos ejemplos de los archivos CSV intermedios más relevantes que se crearon en la carpeta “tablas”, ubicada dentro del directorio del proyecto. A continuación se detallan los resultados de cada una de las etapas.

### 3.3.1. Extracción de datos

Como se ha mencionado anteriormente, el portal de Revistas UM utiliza la plataforma Open Journal System para la publicación de los datos. Tras el análisis del acceso a los datos se observó que cada revista dispone de una URL distinta. Para extraer los datos de todas las revistas de forma automatizada fue necesaria la extracción previa de los puntos de acceso. Para ello, se utilizaron dos técnicas de extracción combinadas con los siguientes resultados:

**Revistas** (archivo revistas.csv): Información general de cada revista científica en formato tabular. Utilizando la técnica *web scraping*, se obtuvieron las URL de acceso a las revistas con un 100% de éxito (43 revistas) y una elevada cobertura de propiedades. Los resultados de la extracción de propiedades en porcentajes fueron: 93% DOI, 100% ISSNE, 53% ISSN, y 100% imágenes (portadas de revistas). En el [Anexo I](#) se presenta un fragmento del script utilizado para la extracción de los registros OAI-PMH en formato XML.

**Artículos** (archivo articulos\_oai\_dc.xml): Información de los registros (artículos) en formato XML. A partir de los puntos de acceso de la extracción anterior, se descargaron los registros en formato XML mediante el verbo “ListRecords” del protocolo OAI-PMH. Este comando proporciona un listado de registros utilizando un sistema de paginación. Para recopilar el contenido completo del repositorio, se implementó un proceso iterativo que recorre todas las páginas hasta obtener la totalidad de los registros disponibles. En este proceso se aplicaron filtros para descartar registros no deseados (por ejemplo, registros marcados como borrados o el filtrado de artículos con la palabra “revisores” en su título, que no resultan relevantes). Los resultados más relevantes de la extracción de registros fueron los siguientes:



1. **Registros extraídos:** 24.650
2. **Tiempo de ejecución:** 23 minutos
3. **Registros eliminados:** Se encontraron 296 registros marcados con status="deleted" (1,2% del total) que fueron omitidos.
4. **Filtrado de contenido no relevante:** Para evitar incluir listas de evaluadores, se omitieron 40 registros (0,16%) por contener la palabra "revisores" en el título.
5. **Manejo de errores:** El sistema de reintentos (3 por solicitud) y pausas de 1 segundo entre paginaciones permitió descargar el 98,4% de los registros del repositorio, evitando que el script se detuviera en los errores de lectura de registros.

Este método combinado de extracción permitió garantizar la correcta vinculación de los artículos a su revista, manteniendo los identificadores únicos OAI-PMH (<identifier>) y por tanto la procedencia de cada registro.

### 3.3.2. Limpieza y normalización

En los procesos de limpieza y normalización se utilizaron expresiones regulares (véase glosario) para eliminar caracteres no deseados como espacios dobles, saltos de línea y otros similares. Durante la limpieza, los datos se organizaron en diferentes archivos CSV intermedios directamente relacionados con las entidades del conjunto de datos:

**Números/volúmenes** (archivos `artículos_volumenes.csv` y `números.csv`): Su cometido es vincular cada artículo con el volumen donde fue publicado, creando la estructura jerárquica Revista → Volumen → Artículo. Esto se consiguió a partir del procesamiento del metadato <dc:source>, donde se emplearon expresiones regulares para normalizar y separar cadenas con datos agregados. Se identificaron un total de 1.596 volúmenes únicos en el conjunto de datos. Para validar el proceso, se contrastaron los resultados con los publicados en el portal web de Revistas UM, realizando un muestreo visual.

Los resultados obtenidos mostraron una cobertura del 100% en el elemento <dc:source>, por lo que todos los artículos quedaron vinculados con su respectivo volumen. Así mismo, también se obtuvo un 100% de cobertura en la vinculación entre volúmenes y revistas, con una asignación que oscila entre 7 y 102 volúmenes por revista. Por ejemplo, en revistas como Daimon y Red se encontraron más de 100



volúmenes mientras que otras como Azufre no superaron las 15 ediciones. En cuanto a la distribución de artículos por volumen, el informe generado por el script identificó 18 volúmenes con menos de 3 artículos. Se realizó una inspección en el portal web de los volúmenes con mayor y menor cantidad de artículos sobre un muestreo del 2%, que confirmó la consistencia tanto del proceso de normalización como de la calidad del metadato de origen.

**Artículos** (archivo articulos.csv): A partir del archivo XML procedente de la extracción, se creó una tabla de 15 campos con los metadatos más relevantes de los artículos. En caso de metadatos que venían en varios idiomas, se utilizó el atributo `xml:lang` para priorizar los elementos en castellano. Se encontró el atributo de idioma en un 95,5% de los artículos, lo que mejoró el PLN aplicado posteriormente. Es destacable la amplia cobertura conseguida en el identificador DOI, que hubo que procesar como valor repetido del metadato `<dc:source>` empleado también para identificar el volumen. En este caso se consiguieron un total de 18.732 identificadores, lo que supone un 76% de artículos. Por contra, el metadato `<dc:description>` no estaba presente en 4.568 registros, lo que supone el 18% de los artículos sin resumen, con el consiguiente impacto en la clasificación temática posterior.

**Autores** (archivos autores.csv y autores\_normalizados.csv): Extracción de los autores de artículos almacenados en el archivo XML. En el proceso se identificaron 35.109 entradas únicas. La entrada “Autores, Varios” tenía 110 artículos asociados, reflejando la práctica editorial de atribución de obras colaborativas en términos agrupados. En la detección de inconsistencias se encontraron 27 autores en blanco y 44 que comenzaban con caracteres no alfabéticos, lo que evidencia un porcentaje de incidencias inferior al 0.2% del total de autores. Las incidencias detectadas se almacenaron en un fichero auxiliar en formato CSV para facilitar una revisión posterior.

La detección de similitudes aplicada mediante el algoritmo RapidFuzz identificó 80 pares de autores potencialmente duplicados con similitud  $\geq 98\%$ , incluyendo variantes como "González, María" vs "Gonzalez, Maria" (sin tilde). El resultado se almacenó en un fichero csv para su revisión posterior. En la Tabla 4 se incluye una muestra de estos archivos de revisión.



Tabla 4. Archivos de revisión de autores.

Incidencias autores		Pares de autores similares	
Autor	Identificador	Autor 1	Autor 2
Ivarez Pérez, Pablo	oai:revistas.um.es:article/151191	Araujo, MC	Araujo, MC.
-----, -	oai:revistas.um.es:article/92	Leite, JL.	Leite, J. L.
., RIE	oai:revistas.um.es:article/98911	Osma, Jorge	Osma , Jorge

Posteriormente a la detección de similitudes, se aplicó un proceso de normalización que redujo un 5.4% el número de autores, aplicando técnicas de agrupamiento para unificar variantes con tildes o guiones y conservando la variante asociada a un mayor número de artículos. Por ejemplo, se unificó "Francisca García" y "Francisca Garcia" en una sola entrada, y "Martínez-López, A." y "Martínez López, A." en otra. La normalización consiguió fusionar 1.722 grupos de autores similares.

**Palabras clave** (archivo keywords.csv): Extracción de las palabras clave en español de los artículos a partir del archivo XML. El proceso detectó 39.892 términos únicos en el metadato <dc:subject> de los registros. Este elemento tenía consignado el atributo xml:lang, por lo que la extracción asignó prioridad a los términos en castellano. Sin embargo, se comprobó que en algunos casos el valor de este atributo no coincidía con el idioma real del contenido, lo que obligó a asumir un cierto nivel de ruido en los resultados.

La palabra clave con más apariciones fue “enfermería”, asociada a 386 artículos. Para la detección de incidencias y similitudes se utilizaron las mismas técnicas utilizadas en el caso de los autores, identificando 175 incidencias y 178 pares de términos similares. La Tabla 5 contiene una muestra de los archivos de incidencias y similitudes generados para su revisión.

Tabla 5. Archivos de revisión de palabras clave generados en la normalización.

Incidencias palabras clave		Pares de términos similares	
Keyword	Identificador	Keyword 1	Keyword 2
/enfermería	oai:revistas.um.es:article/387251	accón	acción
ħunayn ibn isħaq	oai:revistas.um.es:article/588071	hotel	hotel.
: biodiversity	oai:revistas.um.es:article/409051	útbol	fútbol



En cuanto a las inconsistencias detectadas, se encontraron 60 entradas vacías y 15 términos inválidos, lo que indica una buena gestión del elemento <dc:creator> en el repositorio.

### 3.3.3. Enriquecimiento semántico

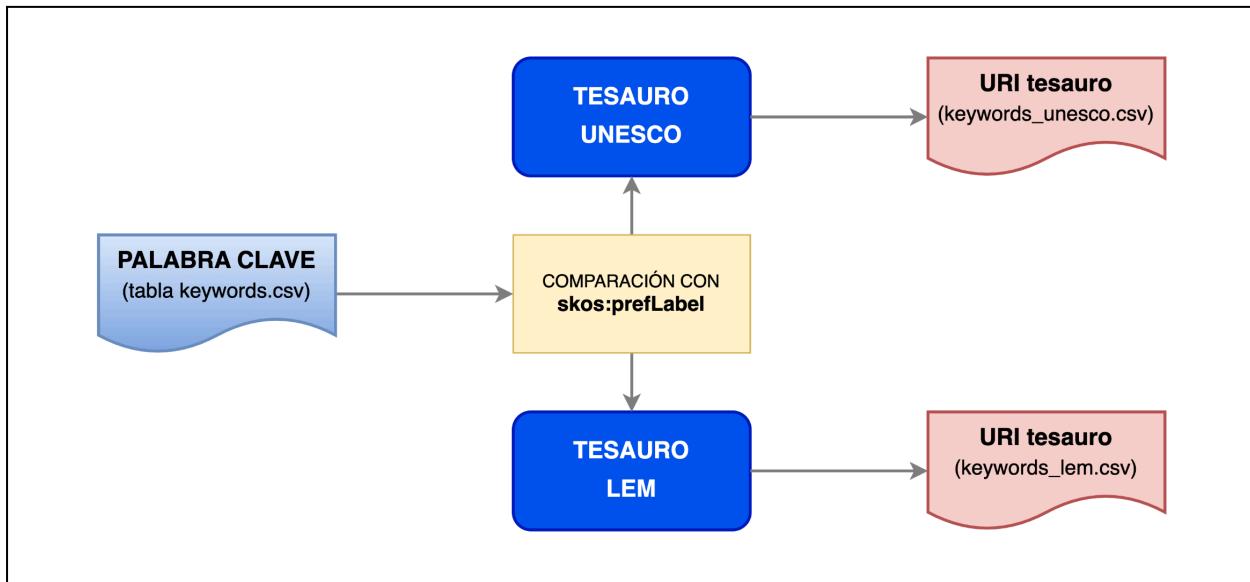
En esta etapa se realizó un enlazado de fuentes externas –proceso denominado reconciliación– para asociar entidades del conjunto de datos. Tras una exploración de las fuentes disponibles, se determinó que las entidades “autor” y “palabra clave” eran las más adecuadas para realizar este proceso. Inicialmente, en la fase de investigación del proyecto, se realizó un intento de vincular los autores con fuentes externas como ORCID o Dialnet. Al no encontrar una metodología que ofreciera unos resultados suficientemente sólidos, el enriquecimiento de autores fue desestimado.

Para las palabras clave, se utilizaron como fuentes externas los tesauros de la UNESCO y la Lista de Encabezamientos de Materia para las Bibliotecas Públicas del Ministerio de Cultura. En el proceso de reconciliación se realizó una búsqueda de correspondencias entre las palabras clave y términos de los tesauros externos. En caso de encontrar equivalencia, se almacenó la URI del término como una propiedad de relación con la fuente externa.

En la Figura 5 se ilustra el proceso de reconciliación de palabras clave del repositorio en el que se realizó una comparación de estas con las etiquetas preferentes (*skos:prefLabel*) de los conceptos en cada tesauro. El resultado de este proceso se guardó en los archivos **keywords\_unesco.csv** y **keywords\_lem.csv**, que contienen la URI (dirección permanente del recurso) del término en la fuente externa.



Figura 5. Diagrama de vinculación con tesauros.



La primera versión del proceso se desarrolló efectuando las consultas directamente en los puntos de acceso SPARQL de los tesauros en línea, lo que llevó un tiempo de ejecución aproximado de 4,5 horas para la consulta de los 39.892 términos. Posteriormente se realizó una redefinición del proceso usando versiones de los tesauros almacenadas localmente. Esta optimización redujo drásticamente el tiempo de procesado, pasando a ejecutarse en menos de 1 minuto.

En el proceso se vincularon 4.588 términos, porcentaje que corresponde a un 11,5% del total. En el tesauro de la Lista de Encabezamientos de Materia para bibliotecas se encontraron 2.987 términos coincidentes (7%), mientras que en el tesauro de la Unesco (véase [Anexo IV](#)) se encontraron 1.874 términos comunes (4,45%).

### 3.3.4. Clasificación temática

La clasificación temática automática representa uno de los aspectos innovadores del sistema desarrollado. Utilizando como referencia las categorías del Tesauro UNESCO, se aplicaron técnicas avanzadas de PLN para asignar automáticamente categorías temáticas a cada artículo, estableciendo una vinculación directa con los términos normalizados de dicho tesauro.

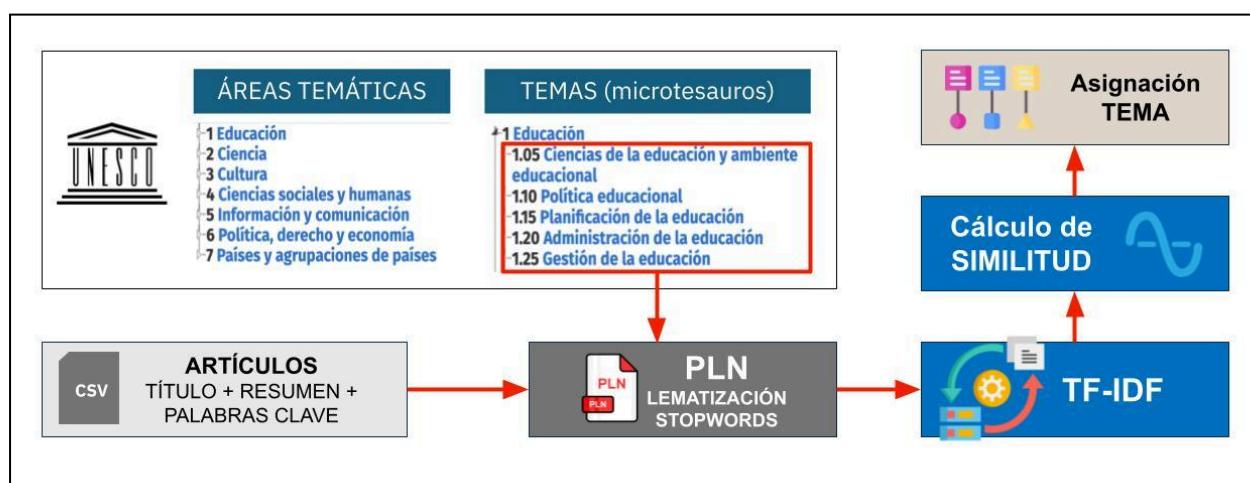
Inicialmente se exploró el uso de técnicas de agrupamiento no supervisado, habituales en disciplinas como la minería de datos. Sin embargo, los resultados preliminares



mostraron inconsistencias significativas en la asignación temática, lo que llevó a replantear el enfoque metodológico.

La estrategia final consistió en adoptar la estructura jerárquica del Tesauro UNESCO, que organiza el conocimiento en 7 áreas temáticas subdivididas en 88 temas o microtesauros. En la Figura 6 muestra el método de clasificación empleado mediante la comparación vectorial del corpus generado para cada artículo con el corpus generado para cada microtesauro (véase [Anexo IV](#)), asignando las categorías mediante la similitud del coseno.

**Figura 6.** Proceso de clasificación temática.



A continuación se detallan los pasos empleados para completar el proceso de clasificación propuesto:

**Corpus de artículos:** Se utilizaron los metadatos de título, resumen y palabras clave de cada artículo, sobre los que se aplicaron técnicas de PLN utilizando las librerías spaCy y NLTK. Este proceso incluyó la lematización (obtener la forma base de las palabras, por ejemplo, cambiar todos los tiempos verbales por el infinitivo) y la eliminación de palabras vacías (artículos y otras entidades gramaticales con escaso valor semántico).

El análisis evidenció que el corpus inicial, con más de 3 millones de palabras, se redujo un 58%. La longitud media resultante fue de 52 palabras por artículo, mostrando una variación directamente proporcional a la extensión del texto original. El artículo con el



corpus más extenso quedó con 297 términos, mientras que hubo algunos artículos que quedaron sin términos en el corpus. La Tabla 6 presenta una muestra del texto de entrada y la salida obtenida tras el PLN.

**Tabla 6.** Ejemplo del texto de entrada en un registro y su salida tras el PLN.

Texto de entrada		
Título	Resumen	Palabras clave
REVISIÓN TAXONÓMICA Y APORTACIONES COROLÓGICAS PARA EL GÉNERO GLADIOLUS L. (IRIDACEAE) EN LA REGIÓN DE MURCIA	Revisión de los taxones del género Gladiolus citados para la Región de Murcia. Para cada taxón se incluye el nombre correcto, sinónimos, descripción, ecología y corología, así como breves comentarios nomenclaturales. Se realizan correcciones y comentarios sobre material dudoso, y nuevas aportaciones corológicas de las especies menos citadas.	Gladiolus Iridaceae Murcia SE España corología

Texto de salida: Resultado de procesado (corpus)
revisión taxonómico aportación género región revisión taxón género citado región taxón incluir nombre correcto sinónimos descripción ecología breve comentario nomenclatural realizar corrección comentario material dudoso aportación especie citado

La posibilidad de discriminar los metadatos en otros idiomas mediante el atributo `xml:lang` fue fundamental para obtener resultados óptimos. En el conjunto de datos se encontraron 338.280 atributos `<xml:lang>`, de los cuales 188.214 correspondían al idioma castellano. Sin embargo, se detectaron casos en los que el valor del atributo no coincidía con el idioma real del contenido. Para mejorar los resultados, se implementó una estrategia específica que consistió en filtrar los términos mediante el corpus en castellano de la librería NLTK.

**Corpus del tesoro:** Para obtener el corpus del tesoro se aplicaron las mismas técnicas utilizadas para obtener el corpus de los artículos. En la Tabla 7 se puede observar un fragmento de un grupo temático. Los datos de área e identificador también se guardaron para construir posteriormente el archivo TTL de grupos temáticos.

**Tabla 7.** Resultado de la técnica PLN en el corpus de un microtesauro.

Tema en el Tesauro		
Área: Educación	ID microtesauro: 1.05	Grupo temático: Ciencias de la educación y ambiente educacional
Términos que componen el microtesauro procesados (fragmento)		
psicología educación psicología educativo psicopedagogía psicología adolescente psicología niño pedagógico rendimiento escolar psicología desarrollo proceso aprendizaje psicosociología educación aprendizaje actitud docente clase actitud estudiante participación profesor actitud estudiante actitud estudiantil [...]		



**Cálculo de frecuencia de términos:** Mediante la técnica TF-IDF se realizó un cálculo de frecuencia de aparición de términos. Esta técnica calcula la frecuencia de aparición de términos en cada artículo en relación con el total del corpus de artículos. Este cálculo convierte los términos en vectores que podrán ser comparados posteriormente, además de proporcionar estadísticas sobre los términos más frecuentes. Se procesaron 23.520 artículos, con un total de 32.602 palabras únicas. La palabra más frecuente fue “estudio” con 11.781 apariciones. Como resultado, se generó una nube de palabras (Figura 7) que permite visualizar la distribución de los términos en el corpus. El tamaño de cada palabra representa proporcionalmente su número de apariciones.

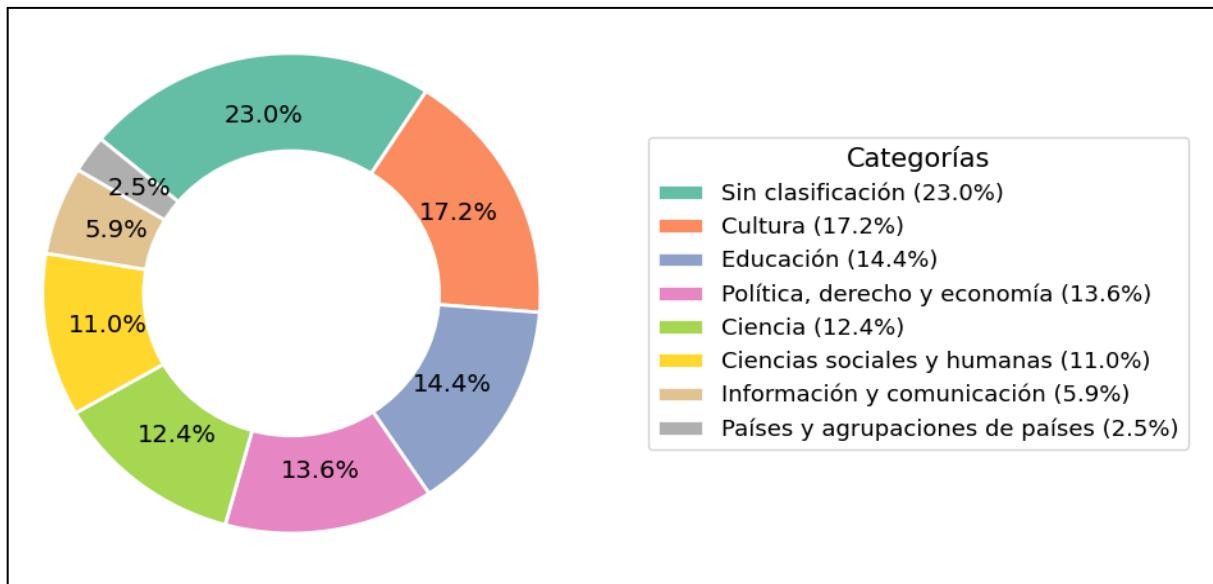
**Figura 7.** Nube de palabras con la frecuencia de términos en el corpus artículos.



**Cálculo de similitud del coseno:** Utilizando los corpus anteriores, se calculó la similitud de los artículos con los tema del tesauro. Esta técnica compara los vectores TF-IDF obtenidos anteriormente para asignar a cada artículo el mejor puntuado en el cálculo. El promedio de similitud se situó en torno al 1%. Para asegurar una mayor coherencia en la clasificación, se utilizó un umbral de similitud del 5%. El resultado fue la clasificación de 18.985 artículos, un 72% del total. En la Figura 8 se puede observar la distribución temática del conjunto de revistas. El 23% de artículos no clasificados se corresponde principalmente a los 4.568 artículos (18%) que no contaba con resumen, por lo tanto, no disponían de texto suficiente para calcular la clasificación.



**Figura 8.** Distribución temática de las revistas en el tesoro de la UNESCO.



Para comprobar la coherencia en la clasificación, se almacenó el término clave que generó la asignación entre los dos corpus en el archivo corpus.csv. Tras realizar una inspección visual (véase [Anexo IV](#)) se observó una precisión aceptable en las clasificaciones asignadas.

### 3.3.5. Generación de ficheros RDF

En la última etapa del proceso de transformación se siguió el perfil de aplicación RDF para construir el conjunto de datos final. A diferencia del prototipo creado en la investigación inicial, donde se utilizó la herramienta de transformación de datos OpenRefine, la creación de los ficheros se realizó mediante la librería Rdflib. Este proceso se subdividió en dos pasos, simplificando así la creación del conjunto de datos final. En primer lugar, se generó un archivo en formato TTL para cada entidad, unificando finalmente todos los ficheros en un único archivo con el dataset completo.

En el primer paso del proceso, se usaron los datos de los archivos CSV intermedios para crear un archivo TTL por cada entidad, generando identificadores de recurso únicos conforme al espacio de nombres. En la Tabla 8 se puede observar la correspondencia entre los archivos TTL y los archivos CSV, así como los resultados cuantitativos de entidades y tripletes generadas:

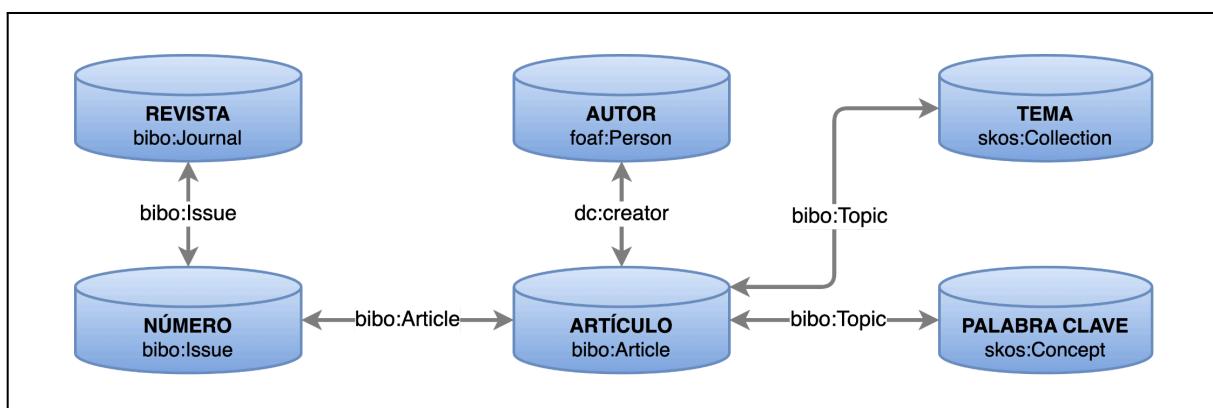


Tabla 8. Resultados de la creación del conjunto de datos.

Entidad	Archivo CSV intermedio	Nº entidades	Nº de triplets
Revistas	revistas.csv	43	321
Números / Volúmenes	articulos_volumenes.csv números.csv	1.596	31.034
Artículos	articulos.csv   clusters.csv	24.650	257.195
Autores	autores.csv	33.213	148.603
Palabras clave	keywords.csv	39.892	164.031
> Enriquecidas con Tesauro LEM	keywords_lem.csv	2.814	
> Enriquecidas con Tesauro UNESCO	keywords_unesco.csv	1.774	
> Enriquecidas con ambos tesauros	-	727	
Grupos temáticos	tesauro_unesco_corpus.csv	88	533
<b>Conjunto de datos final</b>		<b>99.483</b>	<b>781.027</b>

Una vez obtenidos los ficheros TTL individuales, en el segundo paso de esta etapa se consolidaron todos los componentes en un único grafo, añadiendo relaciones inversas en todas las entidades. Por ejemplo, a los artículos vinculados con un autor mediante la propiedad “tieneAutor”, se les añadió la relación inversa “esAutorDe” desde el autor hacia el artículo. En la Figura 9 se ha realizado una representación simplificada del grafo RDF resultante, que muestra las principales entidades y sus relaciones según el perfil de aplicación.

Figura 9. Diagrama del modelo RDF aplicado al conjunto de datos unificado.



Finalmente, se añadieron metadatos identificativos al conjunto de datos utilizando vocabularios estándar como DCAT y DCTERMS, incluyendo información sobre el título, descripción, autoría, licencia y fecha de creación.



El tiempo total de ejecución fue de 6 minutos y 51 segundos. El dataset unificado en formato TTL ocupó 67,3 Mb, lo que representa una reducción del 28% respecto al archivo XML original (93,5 Mb). En el [Anexo V](#) se adjunta un fragmento del dataset en formato TTL. El dataset completo se encuentra disponible en el repositorio [Github](#) del proyecto.

### 3.4. Interfaz web de búsqueda y consulta

Para la explotación del dataset, la aplicación desarrollada se basó en una interfaz web para la presentación dinámica de los datos. Su estructura jerárquica permite explorar la información de manera progresiva, facilitando el descubrimiento de las relaciones semánticas entre autores, artículos y temas. El desarrollo se realizó aplicando el modelo de tres capas y un despliegue modular. Además, se incorporaron técnicas de visualización progresiva y sistemas de filtrado avanzados para mejorar la experiencia de búsqueda y navegación.

#### 3.4.1. Arquitectura de la aplicación

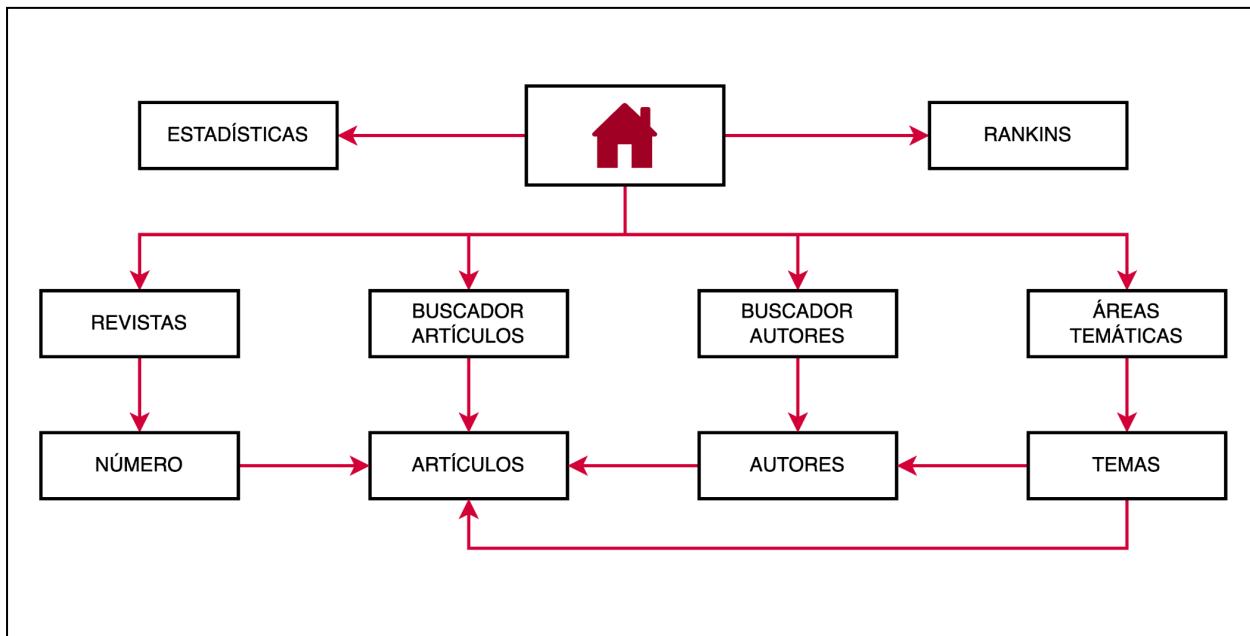
El diseño de la aplicación siguió una arquitectura web centrada en el usuario, implementando un sistema de navegación contextual que permite transitar desde vistas agregadas (listados generales) hasta detalles específicos de cada entidad, manteniendo siempre visibles las relaciones semánticas. La navegación se dividió en tres niveles jerárquicos:

- Vista de resumen: Estadísticas y rankings.
- Vistas de listado: Presentación filtrada de entidades.
- Vistas de detalle: Información completa con relaciones asociativas y navegación transversal.

La Figura 10 muestra la arquitectura general desarrollada. Aunque el flujo indica una jerarquía, el enlazado semántico presente en todos los listados permite la navegación no lineal por todo el contenido.



Figura 10. Diagrama de la arquitectura de la aplicación web.



Cada elemento interactivo (enlaces, botones de filtrado, tablas ordenables) se configuró para activar consultas SPARQL dinámicas que actualizan el contenido. Los enlaces semánticos entre entidades (autor → artículos → revista) permiten descubrir conexiones implícitas en los datos, implementando el principio de navegación por asociación. En la presentación de información se utilizaron técnicas de visualización progresiva:

- Tarjetas resumen: Muestran recuentos globales con enlaces a listados completos.
- Rankings temáticos: Destacan las entidades más relevantes mediante algoritmos de ponderación (frecuencia de aparición, coautorías).
- Perfiles enriquecidos: Combinan metadatos básicos con análisis derivados (redes de colaboración de autores, distribución temática).

Los sistemas de filtrado se implementaron mediante consultas SPARQL dinámicas en las que se combinan múltiples criterios (título, autor, revista, etc.), permitiendo añadir la cantidad de criterios necesaria para construir la estrategia de búsqueda deseada. La interfaz oculta la complejidad técnica tras formularios intuitivos que generan consultas parametrizadas automáticamente.



La aplicación se desarrolló en PHP bajo una estructura modular basada en el sistema cliente-servidor y arquitectura de tres capas lógicas (presentación, aplicación y acceso a datos). La capa de aplicación implementa la lógica de interacción con el usuario, utilizando código HTML y hojas de estilo CSS para generar la capa de presentación. En la capa de acceso a datos se implementaron algoritmos para la generación dinámica de consultas SPARQL, utilizando un servidor Apache Jena Fuseki donde previamente se subió el conjunto de datos obtenido en el proceso de transformación (Figura 11).

Figura 11. Interfaz web del servidor Apache Jena Fuseki.

The screenshot shows the Apache Jena Fuseki interface at the URL [/datarevistas](#). At the top, there is a navigation bar with links for datasets, manage, and help. Below the navigation, there are buttons for query, add data, edit, and info. The main area is titled "SPARQL Query" and contains a text input field for entering SPARQL queries. The input field contains the following query:

```
6 SELECT ?GrupoTemaArea (COUNT(?Articulo) AS ?numArticulos)
7 WHERE {
8     ?Articulo a ontorevistas:Articulo ;
9         ontorevistas:perteneceAGrupoTema ?GrupoTema .
10
11     ?GrupoTema ontorevistas:grupoTemaArea ?GrupoTemaArea .
12 }
13 GROUP BY ?GrupoTemaArea
14 ORDER BY DESC(?numArticulos)
```

Below the query input, there are tabs for "Selection of triples" (selected) and "Selection of classes". To the right, there are buttons for "Prefixes" (rdf, rdfs, owl, xsd) and "Content Type (SELECT)" (JSON). The results of the query are displayed in a table with two columns: "GrupoTemaArea" and "numArticulos". The table shows two rows of data:

GrupoTemaArea	numArticulos
1No clasificado	"5665"^^< <a href="http://www.w3.org/2001/XMLSchema#integer">http://www.w3.org/2001/XMLSchema#integer</a>
2Cultura	"4234"^^< <a href="http://www.w3.org/2001/XMLSchema#integer">http://www.w3.org/2001/XMLSchema#integer</a>

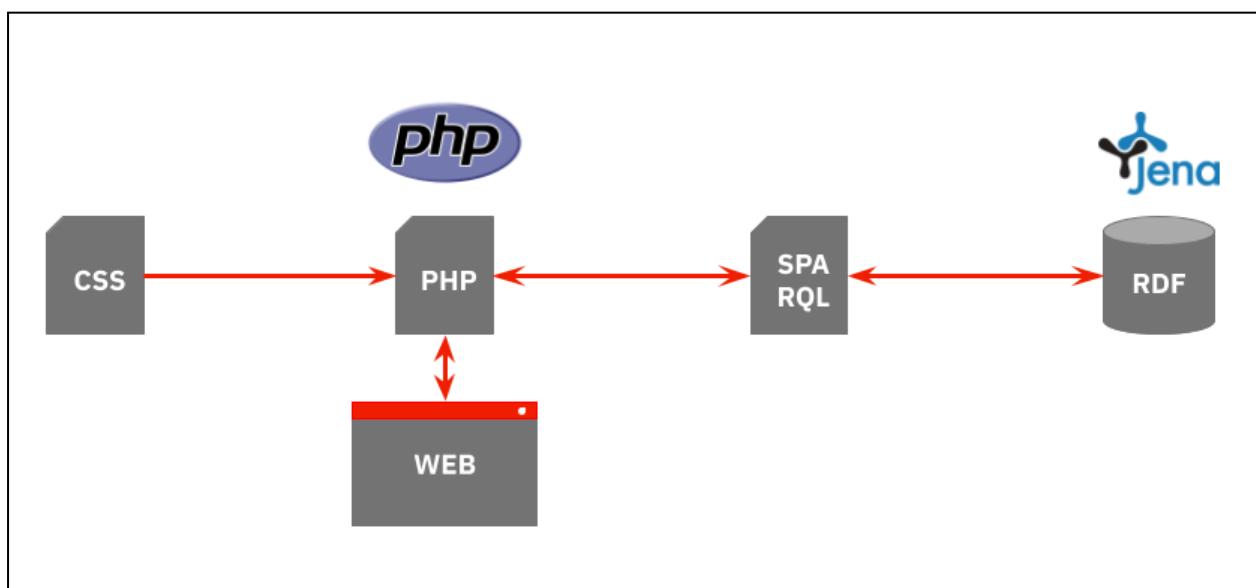
Siguiendo la arquitectura de la información planificada (véase Figura 10), se crearon módulos independientes para cada entidad principal (revistas, artículos y autores). Cada módulo se dividió en tres archivos con la siguiente estructura:



- Un archivo PHP que contiene la lógica de la aplicación, gestionando a su vez las capas de visualización y acceso a datos.
- Un archivo de consultas SPARQL específico para cada módulo que genera las consultas dinámicas necesarias para cada elemento.
- Una hoja de estilos CSS para ajustar la visualización de elementos del módulo. Esto facilitó mantener la coherencia visual, usando un diseño de paleta de colores, tipografía y maquetación basado en la imagen corporativa de la Universidad de Murcia.

En la Figura 12 se puede ver la conexión entre los módulos desarrollados. Esta modularidad permite escalar el sistema añadiendo nuevas funcionalidades sin afectar al código existente. La comunicación entre componentes se realiza mediante consultas parametrizadas que se construyen de forma dinámica y recuperan únicamente los datos necesarios para cada vista, optimizando el rendimiento general de la aplicación.

**Figura 12.** Diseño modular de componentes de la aplicación.



### 3.4.2. Desarrollo de la aplicación

A partir de la estructura modular propuesta, se desarrolló la estructura específica de la aplicación, cuyo resultado fue, como se puede observar en la Tabla 9, una serie de módulos que trabajan de manera conjunta para ofrecer una experiencia de navegación completa. Adicionalmente, se crearon archivos auxiliares con funciones comunes para todos los módulos, así como un archivo de configuración global del punto de acceso.



En el [Anexo VI](#) se muestra un fragmento del código PHP de aplicación y su correspondiente módulo asociado para realizar una consulta dinámica. La aplicación está disponible en el repositorio [Github](#) del proyecto.

**Tabla 9.** Resumen de los módulos de la aplicación y ficheros de configuración.

Módulos de aplicación	
<b>index.php</b>	Genera la página de inicio
<b>index_bloques.php</b>	Genera los ranking de la página de inicio
<b>areas-lista.php</b>	Crea el listado de áreas temáticas
<b>areas.php</b>	Crea la lista de temas por cada área
<b>tema.php</b>	Genera el listado de entidades por cada tema
<b>articulo.php</b>	Genera la ficha completa de los artículos
<b>articulos-buscador.php</b>	Crea la página del buscador de artículos
<b>autor.php</b>	Genera la lista de autores
<b>autores-buscador.php</b>	Crea la página del buscador de autores
<b>autores-lista.php</b>	Genera la lista de autores
<b>revistas-lista.php</b>	Genera la lista de revista
<b>revista.php</b>	Genera la ficha completa de la revista
<b>numero.php</b>	Genera la ficha completa del número
Módulos con funciones comunes	
<b>menu_principal.php</b>	Genera la cabecera y el menú superior de todas las páginas
<b>sparqlquerydispatcher.php</b>	Realiza las consultas en el punto de acceso
<b>sparql_prefijos.php</b>	Configura los espacios de nombres para las consultas
Archivo de configuración	
<b>endpoint_url.php</b>	Contiene la dirección URL del punto de acceso

Para mostrar el resultado final de la aplicación, a continuación se muestran diferentes capturas de pantalla con los elementos más relevantes desarrollados, donde se puede apreciar la sencillez del diseño de la interfaz centrada en la visualización de los rankings y enlaces entre los elementos.



La página de inicio se divide en dos secciones. En la primera sección (Figura 13) se muestran las estadísticas generales: revistas, artículos, autores y áreas temáticas. Pulsando sobre cualquier de las tarjetas de visualización se accede a secciones con información ampliada.

**Figura 13.** Página de inicio de la aplicación web. Sección de estadísticas.

The screenshot shows the homepage of the 'Revistas UM' application. At the top, there is a red header bar with the logo 'edit.um', the title 'Revistas UM', and the University of Murcia logo. Below the header, a section titled 'Explorador Semántico de Revistas Científicas UM' provides general information about the repository. It states that it is based on the OAI-PMH protocol of the University of Murcia (2025) and is a semantic search tool for scientific journals. It mentions that the system uses RDF modeling and allows exploring semantic relations between journals, articles, authors, and UNESCO topics. The main content area is divided into sections: 'Contenido' (Content) which displays statistics for journals (43), articles (24650), and authors (33212); and 'Áreas temáticas (basadas en el Tesauro de la UNESCO)' (Thematic Areas based on the UNESCO Thesaurus) which lists various categories with their article counts. The 'Contenido' section has three cards:

Revistas	Artículos	Autores
43	24650	33212

The 'Áreas temáticas' section has eight cards:

No clasificado	Cultura	Educación	Política, derecho y economía
5665 artículos	4234 artículos	3553 artículos	3353 artículos
Ciencia	Ciencias sociales y humanas	Información y comunicación	Países y agrupaciones de países
3058 artículos	2705 artículos	1464 artículos	618 artículos



## Transformación de metadatos OAI-PMH a conjuntos de datos semánticos RDF en revistas científicas y repositorios institucionales

La segunda parte de la página de inicio titulada “Destacados” (Figura 14) se compone de varias tablas con diferentes rankings: revistas con más artículos, autores con más artículos, etc. Todas las filas que aparecen en las tablas son enlaces que llevan a las páginas de detalles del elemento en cuestión, permitiendo comenzar la exploración desde cualquier elemento.

**Figura 14.** Página de inicio de la aplicación web. Sección de rankings.

Destacados								
Top 10 Revistas con más Artículos			Top 10 Revistas con más Autores		Top 10 Autores con más Artículos			
1	Daimon Revista Internacional de Filosofía	1905	1	Enfermería Global	5502	1	Autores, Varios	192
2	Enfermería Global	1786	2	Anales de Psicología / Annals of Psychology	4318	2	González Blanco, Antonino	91
3	Anales de Psicología / Annals of Psychology	1763	3	Cuadernos de Psicología del Deporte	2521	3	Molina Gómez, José Antonio	74
4	Cuadernos de Psicología del Deporte	1016	4	SPORT TK-Revista EuroAmericana de Ciencias del Deporte	2080	4	Educatio Siglo XXI, Revista	62
5	Antigüedad y Cristianismo	1013	5	Revista de Investigación Educativa	1706	5	Hernández Mendo, Antonio	59
6	Educatio Siglo XXI	966	6	Daimon Revista Internacional de Filosofía	1357	6	Daimon, Revista	56
7	Revista de Investigación Educativa	945	7	Anales de Biología	1337	7	Zapata Ros, Miguel	51
8	Monteagudo. Revista de Literatura Española, Hispanoamericana y Teoría de la Literatura	863	8	Revista Electrónica Interuniversitaria de Formación del Profesorado	1236	8	Egea Vivancos, Alejandro	45
9	Myrtia	806	9	Revista de Educación a Distancia (RED)	1223	9	White, Heather	42
10	Cuadernos de Turismo	782	10	Educatio Siglo XXI	1220	10	Cebrián Abellán, Aurelio	36

Top 10 Artículos con más autores						
#	Revista	Número Volumen	Título	Número de Autores	Fecha	Tema
1	Anales de Biología	Anales de Biología; Núm. 43 (2021)	Estado de conservación de las tortugas marinas en España (revisión del periodo 2013-2018)	20	2021-12-15	Geografía y oceanografía
2	Anales de Psicología / Annals of Psychology	Anales de Psicología / Annals of Psychology; Vol. 33 Núm. 3 (2017)	Propiedades psicométricas de la Liverpool Stoicism Scale (LSS) en una cohorte de pacientes con cáncer resecado en tratamiento adyuvante.	18	2017-07-21	Patología
3	Anales de Biología	Anales de Biología; Núm. 31 (2009)	Aproximación a la checklist de los gipsófitos ibéricos	16	2009-12-01	Ciencias naturales
4	Anales de Veterinaria de Murcia	Anales de Veterinaria de Murcia; Vol. 28 (2012)	Comparative value of microscopy, serology and real time pcr in the diagnosis of asymptomatic canine Leishmania infantum infection	15	2012-12-01	Evaluación de la educación
5	Revista de Educación a Distancia (RED)	Revista de Educación a Distancia (RED); Vol. 24 Núm. 78 (2024); IA generativa, ChatGPT y Educación. Consecuencias para el Aprendizaje Inteligente y la Evaluación Educativa.	Máquina contra máquina: Modelos de Lenguaje de Gran Escala (LLM) en Exámenes de Alto Riesgo de Aprendizaje Automático Aplicado con apuntes	15	2024-05-30	Evaluación de la educación
6	Anales de Biología	Anales de Biología; Núm. 39 (2017)	Catálogo de las aves de la Región de Murcia (España)	14	2017-01-23	No clasificado
7	Cuadernos de Psicología del Deporte	Cuadernos de Psicología del Deporte; Vol. 17 Núm. 3 (2017)	Relación entre la grasa corporal y la expresión de ira en personas que realizan ejercicio regularmente	14	2017-12-10	Psicología
8	Enfermería Global	Enfermería Global; Vol. 3 Núm. 2 (2004): #5 - Noviembre	ANÁLISIS DE LA NUEVA PROPUESTA EDUCATIVA PARA LA EDUCACIÓN SUPERIOR: OPINIÓN SOBRE EL DOCUMENTO DE CONVERGENCIA EUROPEA DESDE LA PERSPECTIVA DEL ALUMNADO DE LA DIPLOMATURA DE ENFERMERÍA.	14	2004-11-01	Sistemas y niveles de enseñanza
9	Myrtia	Myrtia; Vol. 23 (2008)	RESEÑAS	14	2008-12-01	Literatura
10	Anales de Psicología / Annals of Psychology	Anales de Psicología / Annals of Psychology; Vol. 39 Núm. 2 (2023); mayo - septiembre	Contribución de la participación en actividades significativas sobre la salud mental en población española durante el confinamiento por COVID-19	13	2023-04-27	Patología

Pedro Otálora. TFG. Conjunto de datos RDF con información de revistas científicas de la Universidad de Murcia creado a partir de datos OAI-PMH.



## Transformación de metadatos OAI-PMH a conjuntos de datos semánticos RDF en revistas científicas y repositorios institucionales

La página correspondiente a la ficha de artículo se divide en dos secciones. La primera sección (Figura 15) muestra los datos detallados del elemento. Todos los elementos están enlazados con la información correspondiente a cada uno. Los enlaces externos permiten acceder al documento original en el repositorio de Revistas UM.

**Figura 15.** Página de ficha de artículo. Sección de detalles.

**Revistas UM**

edit.um

UNIVERSIDAD DE MURCIA

Inicio Revistas Artículos Autores Áreas Temáticas

**Artículo**

**Calidad asistencial percibida y satisfacción de las personas sordas con la atención primaria de un Área de Salud de la Región de Murcia**

**REVISTA DIGITAL ENFERMERÍA GLOBAL**  
<http://revistas.um.es/eglobal>

Revista:	Enfermería Global
Volumen:	Enfermería Global; Vol. 18 Núm. 2 (2019): #54 Abril
Autores:	Conesa Guillén, María de los Ángeles, Pastor Bravo, María del Mar, Cayuela Fuentes, Pedro Simón
Palabras clave:	deficiencia auditiva., sordera, satisfacción usuarios, calidad asistencial, atención primaria
Fecha de Publicación:	2019-02-28
Idioma:	spa   eng
Ref. Bibliográfica:	No especificada
Editor:	Ediciones de la Universidad de Murcia (Editum)
Tema:	Ciencias médicas
DOI:	<a href="https://doi.org/10.6018/eglobal.18.2.344761">https://doi.org/10.6018/eglobal.18.2.344761</a>
Web del Artículo:	<a href="https://revistas.um.es/eglobal/article/view/344761">https://revistas.um.es/eglobal/article/view/344761</a>
Texto completo:	<a href="https://revistas.um.es/eglobal/article/view/344761/258351">https://revistas.um.es/eglobal/article/view/344761/258351</a>

**Resumen**

Objetivo: Describir la calidad asistencial percibida y la satisfacción frente a los servicios de Atención Primaria del Área de Salud II Cartagena del Servicio Murciano de Salud por parte de las personas sordas de Cartagena y comarca. Método: Estudio observacional, descriptivo y transversal. Los datos se recogieron mediante la traducción simultánea a la lengua de signos española del Cuestionario de Evaluación y Mejora de la Calidad Asistencial Global Percibida en Atención Primaria. Se analizaron las variables: edad, sexo, nivel de estudios, tipo de sordera, primera lengua y uso, sistemas o apoyos comunicativos, calidad de atención percibida, percepción de la profesionalidad y trato humano por parte del profesional de la medicina, enfermería y administración y la satisfacción global percibida con su Centro de Atención Primaria. Resultados: La profesionalidad y trato humano recibido por parte del personal médico y administrativo fue percibido como deficiente, considerándose bueno en caso de las enfermeras. La satisfacción global es menor a la de la población general. Existen diferencias estadísticamente significativas entre el tipo de sordera y la profesionalidad percibida, el trato humano y la profesionalidad percibida y entre el sistema o apoyo comunicativo y la calidad de la atención percibida. Conclusiones: Es necesario adaptar la atención en salud que se presta a este colectivo con necesidades especiales a fin de que perciban una atención sanitaria de calidad que derive en un mayor acceso y seguimiento de personas sordas en el sistema sanitario.

**Más artículos del tema Ciencias médicas (235)**

Revista	Número	Título	Autores	Fecha
Anales de Psicología / Annals of Psychology	Anales de Psicología / Annals of Psychology; Vol. 9 Núm. 1 (1993)	MODELOS TEÓRICOS DE PREVENCIÓN EN TOXICOMANÍAS: UNA PROPUESTA DE CLASIFICACIÓN	Carlos Pastor, Juan, López-Latorre, M <sup>a</sup> Jesús	No especificada
Anales de Psicología / Annals of Psychology	Anales de Psicología / Annals of Psychology; Vol. 2 (1985)	LA CUESTIÓN DE LAS DROGAS: CONSIDERACIONES ELEMENTALES	Coy Ferrer, Ernesto	No especificada
Anales de Psicología / Annals of Psychology	Anales de Psicología / Annals of Psychology; Vol. 10 Núm. 2 (1994)	¿CÓMO INFUYE EL CONTROL PERCIBIDO EN EL IMPACTO QUE TIENEN LAS EMOCIONES SOBRE LA SALUD?	Edo Izquierdo, Silvia, Fernández Castro, Jordi	No especificada
Anales de Derecho	Anales de Derecho; Vol. 6 (1984)	La responsabilidad civil de los médicos	Ataz López, Joaquín	1984-12-01
Myrtia	Myrtia; Vol. 6 (1991)	MEDICINA Y PENSAMIENTO EN EL CORPUS HIPPOCRATICUM	López Salvá, Mercedes	1991-12-01
Myrtia	Myrtia; Vol. 6 (1991)	ALGUNAS CARÁCTERISTICAS LINGÜÍSTICAS DE LA OBRA MÉDICA DE ARETEO DE CAPADOCIA	Pérez Molina, Miguel E.	1991-12-01



La segunda sección de la página de ficha de artículo (Figura 16) muestra un listado de artículos de la misma clasificación temática que el artículo visualizado. La sección inferior incluye un navegador de páginas con opciones para cambiar el número de resultados y moverse entre los distintos bloques. Las tablas de resultados se han elaborado manteniendo la coherencia visual y funcional en cabeceras, enlaces, botones, paginación, etc.

**Figura 16.** Página de ficha de artículo. Sección “artículos de la misma categoría”.

Más artículos del tema Ciencias médicas (235)

Revista	Número	Título	Autores	Fecha
Anales de Psicología / Annals of Psychology	Anales de Psicología / Annals of Psychology; Vol. 9 Núm. 1 (1993)	MODELOS TEÓRICOS DE PREVENCIÓN EN TOXICOMANÍAS: UNA PROPUESTA DE CLASIFICACIÓN	Carlos Pastor, Juan, López-Latorre, M <sup>a</sup> Jesús	No especificada
Anales de Psicología / Annals of Psychology	Anales de Psicología / Annals of Psychology; Vol. 2 (1985)	LA CUESTIÓN DE LAS DROGAS: CONSIDERACIONES ELEMENTALES	Coy Ferrer, Ernesto	No especificada
Anales de Psicología / Annals of Psychology	Anales de Psicología / Annals of Psychology; Vol. 10 Núm. 2 (1994)	¿CÓMO INFLUYE EL CONTROL PERCIBIDO EN EL IMPACTO QUE TIENEN LAS EMOCIONES SOBRE LA SALUD?	Edo Izquierdo, Silvia, Fernández Castro, Jordi	No especificada
Anales de Derecho	Anales de Derecho; Vol. 6 (1984)	La responsabilidad civil de los médicos	Ataz López, Joaquín	1984-12-01
Myrtia	Myrtia; Vol. 6 (1991)	MEDICINA Y PENSAMIENTO EN EL CORPUS HIPPOCRATICUM	López Salvá, Mercedes	1991-12-01
Myrtia	Myrtia; Vol. 6 (1991)	ALGUNAS CARÁCTERISTICAS LINGÜÍSTICAS DE LA OBRA MÉDICA DE ARETEO DE CAPADOCIA	Pérez Molina, Miguel E.	1991-12-01
Papeles de Geografía	Papeles de Geografía; Núm. 20 (1994)	INFLUENCIA DE LOS ELEMENTOS Y FACTORES GEOGRÁFICOS EN LA EPIDEMIOLOGÍA DE LA BRUCELOSIDAD DEL GANADO OVINO Y CAPRINO	Crespo León, Fernando	1994-12-01
Daimon Revista Internacional de Filosofía	Daimon Revista Internacional de Filosofía; Núm. 11 (1995)	UN SOLO SEXO. INVENCIÓN DE LA MONOSEXUALIDAD Y EXPULSIÓN DEL HERMAFRODISMO (ESPAÑA, SIGLOS XV-XIX)	Moreno Mengíbar, Andrés, Vázquez García, Francisco	1995-12-01
Revista Murciana de Antropología	Revista Murciana de Antropología; Núm. 3 (1996). Congreso Creencias y mitos: su papel en la configuración del sistema socio-cultural 2	UNA APORTACIÓN AL ESTUDIO DE LA MEDICINA POPULAR EN LA REGIÓN DE MURCIA: LAS RECETAS Y REMEDIOS CASEROS EXPERIMENTADOS DEL DR. FR. MIGUEL TENDERO	Gonzalez Castaño, Juan	1996-12-01
Anales de Derecho	Anales de Derecho; Vol. 15 (1997)	La objeción de conciencia en el ejercicio de la medicina	López Hernández, José	1997-12-01

Resultados por página: 10

[1](#) [2](#) [3](#) [4](#) [5](#) [Último >](#) Página 1 de 24

Pedro Otálora. TFG. Conjunto de datos RDF con información de revistas científicas de la Universidad de Murcia creado a partir de datos OAI-PMH.

En las páginas “ficha de revista” y “ficha de número” se han incorporado tablas de ranking de autores, palabras clave y temas, calculados sobre la revista o el número visualizado. También incluyen una tabla inferior donde aparecen todos los artículos que forman parte de la revista, o en su caso, del número que se está visualizando. Las fichas de número incorporan botones de navegación para recorrer los distintos números que componen la revista (Figura 17).



## Transformación de metadatos OAI-PMH a conjuntos de datos semánticos RDF en revistas científicas y repositorios institucionales

Figura 17. Página de ficha de número.

**Revistas UM**

edit.um

Inicio Revistas Artículos Autores Áreas Temáticas

UNIVERSIDAD DE MURCIA

Número/Volumen

Anales de Documentación; Vol. 27 (2024)

**ANALES DE DOCUMENTACIÓN**

Revista de Bibliotecología y Documentación

edit.um

Anales de Documentación

Volumen: Anales de Documentación; Vol. 27 (2024)

Total de Artículos: 8

Total de Autores: 16

Número anterior: Anales de Docum...

Número siguiente: Anales de Docum...

Volver a la revista

**Top 5 Autores**

1 Aguilera Iniesta, Ainhoa	1
2 Andrade da Fonseca, Luciana Di Paula	1
3 Arroyas Serrano, Magín	1
4 Artigas, Wileidys	1
5 Bahia, Eliana Maria dos Santos	1

**Top 5 Palabras Clave**

1 acceso a datos	1
2 acuerdos capitulares	1
3 adhesión léxica	1
4 amazonía brasileña	1
5 archivos eclesiásticos	1

**Top 5 Temas**

1 Sistemas de información documental	2
2 No clasificado	1
3 Planes de estudios	1
4 Enfoque científico	1
5 Administración de la ciencia y de la investigación	1

Artículos Autores

Artículos del número Anales de Documentación; Vol. 27 (2024)

Título del Artículo	Autores	Fecha	Tema
Política de indexación para organizar y representar el conocimiento: estudio de caso en un sistema de bibliotecas de la región amazónica brasileña	Andrade da Fonseca, Luciana Di Paula, Franciele Marques Redigolo	2024-04-03	Sistemas de información documental
Medición de las competencias digitales en Europa y España: una revisión crítica	Aguilera Iniesta, Ainhoa, Vera-Baceta, Miguel-Angel	2024-04-03	Enfoque científico
La producción española de cómics: evolución y tendencias	Osca-Lluch, Julia	2024-10-24	Fuentes de información
Difusión digital en archivos judiciales: análisis de portales de tribunales de justicia brasileros	Rabelo, Leiliane, de Araújo, Paula Carina	2024-10-30	No clasificado
Uso de la minería de texto en fuentes de información para grupos de productores rurales	Camperos Reyes, Jacquelín Teresa, Gonçalves Sant'Ana, Ricardo Cesar	2024-11-26	Investigación y política de la comunicación
Transformaciones tecnológicas en los currículos de los cursos de formación en Biblioteconomía en el Mercosur	Bahia, Eliana Maria dos Santos, Lira, Edna Karina da Silva	2024-11-27	Planes de estudios
Reagrupando una información dispersa en una serie archivística única: documentación de actas y acuerdos capitulares en el archivo de la Catedral de Segorbe	Arroyas Serrano, Magín	2024-04-03	Sistemas de información documental
Revistas científicas indexadas en SciELO Colombia, Perú y Ecuador: estudio del contenido difundido en la red social Facebook	Cueva Estrada, Jorge Manuel, Meleán Romero, Rosana, Sumba Nacipucha, Nicolás, Artigas, Wileidys	2024-11-15	Administración de la ciencia y de la investigación

Resultados por página: 10

1 Página 1 de 1

Pedro Otálora. TFG. Conjunto de datos RDF con información de revistas científicas de la Universidad de Murcia creado a partir de datos OAI-PMH.



La página que muestra el listado de revistas se ha diseñado para mostrar cada revista en una tarjeta que incorpora estadísticas de su contenido (Figura 18). En las áreas temáticas se ha adoptado el mismo método, incorporando en este caso el listado de temas bajo las tarjetas de visualización de áreas.

Figura 18. Página de listado de revistas.

The screenshot shows the 'Revistas UM' section of the edit.um website. At the top, there's a red header bar with the 'edit.um' logo, the title 'Revistas UM', and the 'UNIVERSIDAD DE MURCIA' logo. Below the header, the main content area is titled 'Revistas Disponibles' and displays a grid of 12 journal cards, each representing a different publication:

- AZARBE, Revista Internacional de Trabajo Social y Bienestar**  
Número: 13  
Artículos: 180
- Anales de Biología**  
Número: 45  
Artículos: 651
- Anales de Derecho**  
Número: 47  
Artículos: 568
- Anales de Documentación**  
Número: 39  
Artículos: 496
- Anales de Filología Francesa**  
Número: 32  
Artículos: 709
- Anales de Psicología / Annals of Psychology**  
Número: 89  
Artículos: 1763
- Anales de Veterinaria de Murcia**  
Número: 35  
Artículos: 399
- Antigüedad y Cristianismo**  
Número: 30  
Artículos: 1013
- Arte y Políticas de Identidad**  
Número: 31  
Artículos: 471
- Bioderecho.es**  
Número: 20  
Artículos: 128
- Cartaphilus**  
Número: 20  
Artículos: 436
- Cuadernos de Gestión de Información**  
Número: 7  
Artículos: 56

En las páginas de búsquedas para artículos y autores se ha implementado un sistema de filtros por entidades que permite añadir tantos filtros como sean necesarios, creando las consultas dinámicas de forma transparente para el usuario (Figura 19). En las páginas de ficha de autor se ha incorporado una tabla que muestra otros autores que tienen artículos en común con el autor principal. También se muestra una tabla con los



## Transformación de metadatos OAI-PMH a conjuntos de datos semánticos RDF en revistas científicas y repositorios institucionales

artículos del autor (Figura 20). Este formato de visualización agregada permite descubrir relaciones implícitas en los datos.

**Figura 19.** Detalle de la página de búsqueda de artículos y filtros.

### Buscador Avanzado de Artículos

Artículo: documento Autor: Garcia

Buscar

6 artículos coinciden con los filtros seleccionados. (Artículo: documento), (Autor: Garcia)

Revista	Número	Título del Artículo	Autores	Fecha
Anales de Documentación	Anales de Documentación; Vol. 7 (2004)	'El documento real en la época de los Austrias (1516-1700)'. De Lorenzo Cadalso, Pedro Luis.	García Díaz, Isabel	2004-01-01
Anales de Documentación	Anales de Documentación; Vol. 9 (2006)	Acceso y disfrute de libros antiguos y documentos históricos, como un derecho cultural en México	García, Idalia	2006-01-01
Revista Murciana de Antropología	Revista Murciana de Antropología; Núm. 9 (2002); Pensil del Ave María	EL "PENSIL DEL AVE MARÍA" COMO GÉNERO LITERARIO Y COMO DOCUMENTO HISTÓRICO	González-Blanco García, Elena	2002-12-01
Naveg@mérica. Revista electrónica editada por la Asociación Española de Americanistas	Naveg@mérica. Revista electrónica editada por la Asociación Española de Americanistas; Núm. 29 (2022)	La memoria de la independencia de Cuba a través de los egodocumentos.	Bravo García, Eva	2022-10-17
Revista Murciana de Antropología	Revista Murciana de Antropología; Núm. 12 (2005): Actas del I Congreso sobre etnoarqueología del vino	LA VID Y EL VINO EN LOS DOCUMENTOS MUNICIPALES DE BULLAS	García Caballero, José Luis	2005-12-01
Anales de Documentación	Anales de Documentación; Vol. 10 (2007)	Producción científica de las publicaciones españolas referentes al análisis documental formal (ADF) de documentos: 1990-2006.	Claúd García, Adelina	2008-02-12

Resultados por página: 10

1 Página 1 de 1

Pedro Otálora. TFG. Conjunto de datos RDF con información de revistas científicas de la Universidad de Murcia creado a partir de datos OAI-PMH.

**Figura 20.** Detalle de la ficha de autor con tabla de co-autorías.

### Revistas UM

edit.um

Inicio Revistas Artículos Autores Áreas Temáticas

#### Cayuela Fuentes, Pedro Simón

#### Colaboradores

Nombre del Colaborador	Número de Artículos en Común
Conesa Guillén, María de los Ángeles	1
Pastor Bravo, María del Mar	1

#### Artículos del Autor

Revista	Número	Título del Artículo	Autores	Fecha
Enfermería Global	Enfermería Global; Vol. 18 Núm. 2 (2019): #54 Abril	Calidad asistencial percibida y satisfacción de las personas sordas con la atención primaria de un Área de Salud de la Región de Murcia	Cayuela Fuentes, Pedro Simón, Conesa Guillén, María de los Ángeles, Pastor Bravo, María del Mar	2019-02-28

Pedro Otálora. TFG. Conjunto de datos RDF con información de revistas científicas de la Universidad de Murcia creado a partir de datos OAI-PMH.



## 4. Conclusiones

El proyecto ha conseguido completar, de forma automática, el ciclo de transformación de los metadatos de un repositorio OAI-PMH en un conjunto de datos semántico. Los resultados obtenidos demuestran la viabilidad del proceso en su conjunto.

El análisis inicial de las fuentes de datos originales reveló una importante oportunidad de mejora en la normalización de los datos en origen. La calidad del conjunto de datos final está directamente relacionada con la calidad de los datos en origen. En el proceso de transformación se han aplicado muchos subprocesos para corregir aspectos específicos del repositorio utilizado para el caso de estudio. La adaptación del proceso a otros repositorios conlleva un estudio de estos subprocesos. No obstante, las herramientas de auditoría incorporadas permiten conocer el grado de éxito de la transformación.

La extracción adicional realizada mediante técnicas de *web scraping* ofrece tanto ventajas como inconvenientes: mientras enriquece el esquema de datos con nuevas propiedades, también introduce mayor complejidad técnica. Por otra parte, la elección de extraer los metadatos desde el esquema Dublin Core minimiza los cambios de configuración para trabajar con otros repositorios. En este sentido, cada fuente exige modificaciones específicas en el desarrollo, lo que limita la escalabilidad a corto plazo. Se detecta, por tanto, la conveniencia de una interfaz de parametrización para realizar los ajustes necesarios sin tener que modificar el código.

Los resultados del proceso de clasificación temática actuales ya permiten categorizaciones útiles, pese al alto porcentaje de artículos no clasificados (23%). El porcentaje, vinculado con los artículos que no tienen texto en el resumen, está relacionado a su vez con la calidad de los datos en la fuente de origen.

El proceso completo de transformación en el repositorio de Revistas UM requirió aproximadamente 90 minutos. El rendimiento conseguido evidencia la escalabilidad del proceso de transformación en repositorios de mayor tamaño sin necesidad de elevados recursos para su ejecución.



La aplicación web permite realizar búsquedas facetadas y navegar por relaciones semánticas, cumpliendo su objetivo de exploración de datos. Su interfaz intuitiva permite a usuarios sin formación técnica identificar conexiones entre elementos y realizar consultas complejas mediante filtros combinados.

Con carácter general, el diseño modular y adaptable del proceso de transformación resulta prometedor para su aplicación en otros contextos. A pesar de los desafíos técnicos y las limitaciones inherentes a los datos de partida, el trabajo realizado aporta un marco de referencia válido para la transformación automatizada de repositorios OAI-PMH en conjuntos de datos semánticos RDF. Los logros obtenidos trazan un camino para futuras investigaciones en este campo.

Pese a los resultados obtenidos, el proyecto admite una serie de mejoras y ampliaciones que no han sido acometidas por limitaciones en el cronograma y/o alcance del propio trabajo. A continuación se describen algunas de las mejoras detectadas:

- **Vinculación de autores con identificadores externos:** A pesar de haber incorporado en el proceso un identificador propio, se necesita identificar el ORCID o similar de los autores para mejorar la interoperabilidad del conjunto de datos.
- **Búsquedas federadas en tesauros para palabras clave:** La normalización aplicada en las palabras clave no ha sido suficiente para realizar una reconciliación exitosa con los tesauros externos. La mejora de la reconciliación haría posible un sistema de búsquedas federadas sobre términos relacionados para ampliar el rango de resultados.
- **Clasificación temática:** La escasez o ausencia de texto en los artículos influye negativamente en el proceso de clasificación. Usando el texto completo de los recursos del artículo se puede obtener una cantidad de texto suficiente para mejorar la cobertura y precisión del sistema de clasificación.

Por otra parte, las ampliaciones que admite el proyecto son muy diversas. A continuación se detalla brevemente alguna de las posibilidades:



- **Fichero de configuración global:** Una parametrización de variables y procesos susceptibles de ser diferentes entre repositorios (URL, metadatos específicos) permitiría tener un fichero de configuración para transformar distintos repositorios en bloque, de forma automática y utilizando el mismo código para todos ellos.
- **Actualización incremental del conjunto de datos:** Implementar un sistema para extraer solamente los nuevos registros añadidos desde la extracción anterior. Así se mantendrán los conjuntos de datos actualizados lanzando el mismo proceso de forma periódica. Incluso podría permitir incluir varios repositorios en el mismo grafo de conocimiento.
- **Aplicación web:** Ampliar el alcance de la aplicación para que actúe como un portal común que incluya distintos conjuntos de datos de revistas.

La evolución hacia un portal capaz de realizar consultas en un catálogo de datos conformado por los distintos datasets que se hayan generado en la transformación de OAI-PMH a RDF puede dar respuesta a la mencionada escalabilidad del proyecto. En un contexto de aplicación más amplio, se podría tener una red de repositorios configurados para su transformación y actualización automática, actuando como un “agregador semántico” de revistas científicas. Las posibilidades de descubrimiento a este nivel de interoperabilidad puede contribuir a una mayor difusión de la producción científica institucional.



## 5. Conclusions

The project has successfully completed, in an automated manner, the full cycle of transforming the metadata from an OAI-PMH repository into a semantic dataset. The results obtained demonstrate the overall feasibility of the process.

The initial analysis of the original data sources revealed a significant opportunity for improvement in the normalization of source data. The quality of the final dataset is directly related to the quality of the source data. Throughout the transformation process, numerous subprocesses were applied to correct specific aspects of the repository used as a case study. Adapting the process to other repositories entails a study of these subprocesses. Nevertheless, the incorporated auditing tools make it possible to assess the degree of success of the transformation.

The additional extraction carried out through web scraping techniques offers both advantages and drawbacks: while it enriches the data schema with new properties, it also introduces greater technical complexity. On the other hand, the choice to extract metadata from the Dublin Core schema minimizes configuration changes required to work with other repositories. In this regard, each source demands specific modifications during development, which limits short-term scalability. Therefore, the need for a parameterization interface is identified, allowing necessary adjustments to be made without modifying the code.

The current results of the thematic classification process already allow for useful categorizations, despite the high percentage of uncategorized articles (23%). This percentage, linked to articles lacking abstract text, is in turn related to the quality of the data in the source.

The complete transformation process in the Revistas UM repository required approximately 90 minutes. The achieved performance demonstrates the scalability of the transformation process for larger repositories without the need for significant resources.



The web application enables faceted searches and navigation through semantic relationships, fulfilling its data exploration objective. Its intuitive interface allows users without technical training to identify connections between elements and perform complex queries using combined filters.

In general terms, the modular and adaptable design of the transformation process proves promising for its application in other contexts. Despite the technical challenges and inherent limitations of the source data, the work carried out provides a valid reference framework for the automated transformation of OAI-PMH repositories into RDF semantic datasets. The achievements obtained pave the way for future research in this field.

Despite the results achieved, the project allows for a series of improvements and extensions that have not been addressed due to time constraints and/or the scope of the work itself. Some of the identified improvements are described below:

- **Linking authors with external identifiers:** Although a proprietary identifier was incorporated in the process, it is necessary to identify the authors' ORCID or similar to improve the interoperability of the dataset.
- **Federated searches in thesauri for keywords:** The normalization applied to keywords has not been sufficient to achieve successful reconciliation with external thesauri. Improving reconciliation would enable a federated search system over related terms to broaden the range of results.
- **Thematic classification:** The scarcity or absence of text in articles negatively affects the classification process. Using the full text of article resources would provide enough content to improve the coverage and accuracy of the classification system.

On the other hand, the possible extensions of the project are very diverse. Some of the possibilities are briefly outlined below:

- **Global configuration file:** Parameterizing variables and processes that may differ between repositories (URL, specific metadata) would allow for a



configuration file to transform different repositories in bulk, automatically and using the same code for all of them.

- **Incremental dataset updating:** Implementing a system to extract only the new records added since the previous extraction. This would keep datasets up to date by periodically running the same process, and could even allow for the inclusion of multiple repositories in the same knowledge graph.
- **Web application:** Expanding the scope of the application so it acts as a common portal that includes different journal datasets.

The evolution towards a portal capable of querying a data catalog composed of the various datasets generated from the OAI-PMH to RDF transformation could address the aforementioned scalability of the project. In a broader application context, it would be possible to have a network of repositories configured for automatic transformation and updating, acting as a "semantic aggregator" of scientific journals. The discovery possibilities at this level of interoperability could contribute to greater dissemination of institutional scientific output.



## 6. Referencias bibliográficas

- Abadal, E., Castellà, C. O., Abad-García, F., & Melero, R. (2013). Políticas de acceso abierto a la ciencia en las universidades españolas. *Revista Española de Documentación Científica*, 36(2), e007. <https://doi.org/10.3989/redc.2013.2.933>
- Alcober Fuertes, G. (s.f.). *Recolección automática de datos abiertos*. Deloitte España. Recuperado de <https://www.deloitte.com/es/es/services/consulting/blogs/todo-tecnologia/recolección-datos-abiertos-harvesting-unizar.html>
- Aramayo, F. R. (2017). Etiquetado automático mediante Programación de Lenguaje Natural (PNL) y visualización de la información utilizando repositorios web bajo protocolo OAI-PMH. *Difusiones*, 12(12), 96–118. Recuperado de <http://revistas.ucse.edu.ar/ojsucse/index.php/difusiones/article/view/153>
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5), 34-43. Recuperado de <https://www.lassila.org/publications/2001/SciAm.pdf>
- Cals, J. D., i Caralt, J. C., & Viladrosa, R. C. (2018). Ontologías y web semántica. FUOC. *Fundación para la Universitat Oberta de Catalunya*. Recuperado de <https://www.cartagena99.com/recursos/alumnos/temarios/Ontologias%20y%20web%20semantica.pdf>
- Chaves-Fraga, D., Corcho, O., & Ruckhaus, E. (s.f.). Guía práctica para la publicación de datos enlazados. Iniciativa Aporta, Ministerio de Asuntos Económicos y Transformación Digital, Red.es. Recuperado de [https://datos.gob.es/sites/default/files/doc/file/guia-publicacion-datos-enlazados\\_2\\_0.pdf](https://datos.gob.es/sites/default/files/doc/file/guia-publicacion-datos-enlazados_2_0.pdf)
- Corera-Álvarez, E., & Molina-Molina, S. (2016). La edición universitaria de revistas científicas. *Revista Interamericana de Bibliotecología*, 39(3), 277-288. <https://doi.org/10.17533/udea.rib.v39n3a05>



CRUE-TIC. (2016). *Infraestructura semántica basada en el paradigma de datos abiertos para la gestión de investigación de las universidades españolas*. CRUE. Recuperado de <https://tic.crue.org/wp-content/uploads/2016/07/Memoria-proyecto-H%C3%A9rcules.pdf>

CRUE. (s.f.). *Hércules: Semántica de datos de investigación de universidades*. Recuperado el 24 de febrero de 2025, de <https://tic.crue.org/hercules/>

Datos.gob.es. (2024). *Superando Desafíos en la Publicación de Datos con Linked Data Event Streams (LDES)*. Portal de Datos Abiertos del Gobierno de España. Recuperado de <https://datos.gob.es/es/blog/superando-desafios-en-la-publicacion-de-datos-con-linked-data-event-streams-ldes>

Deroy Domínguez, D. (2022). Las revistas científicas y su rol en la difusión del conocimiento científico. *Revista Cubana de Educación Superior*, 41(Supl. 1). Recuperado el 5 de enero de 2025 de <http://ref.scielo.org/vv992v>

DeVito, N. J., Richards, G. C., & Inglesby, P. (2020). How we learnt to stop worrying and love web scraping. *Nature*, 585(7826), 531–532. <https://doi.org/10.1038/d41586-02-02558-0>. Recuperado de [https://www.researchgate.net/publication/344213676\\_How\\_we\\_learnt\\_to\\_stop\\_worrying\\_and\\_love\\_web\\_scraping](https://www.researchgate.net/publication/344213676_How_we_learnt_to_stop_worrying_and_love_web_scraping)

Ferreras-Fernández T. & Merlo-Vega, J. A. (2015). Repositorios de acceso abierto: un nuevo modelo de comunicación científica. La Revista de la Sociedad ORL CLCR en el repositorio Gredos. *Revista. Sociedad Otorrinolaringológica de Castilla y León, Cantabria y La Rioja*. 6(12): 94-113. <http://hdl.handle.net/10366/126908>

Fernández-Quijada, D. (2012). El uso de tesauros para el análisis temático de la producción científica: Apuntes metodológicos desde una experiencia práctica. *BID: textos universitaris de biblioteconomia i documentació* (29), 1-11. Recuperado de <https://bid.ub.edu/29/fernandez2.htm>

Gobierno de Aragón (s.f.). *Datos enlazados. Datos abiertos Aragón Open Data*. Recuperado el 19 de marzo de 2025 de <https://opendata.aragon.es>



Hernández Pina, F., & Maquilón Sánchez, J. J. (2010). Indicadores de calidad de las revistas científicas y sistema de gestión editorial mediante OJS. *Revista de Investigación Educativa*, 28(1), 13–29. Recuperado de <https://revistas.um.es/rie/article/view/109941>

Interoperable Europe. (s.f.). *Linked Data Event Streams (LDES)*. Recuperado de <https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/linked-data-event-streams-ldes>

Martínez Méndez, F. J., Pastor-Sánchez, J. A., & López Carreño, R. (2020). Linked open data en bibliotecas: estado del arte. *Information Research*, 25(2). Recuperado de <http://InformationR.net/ir/25-2/paper862.html>

Melero, R. (2008). El paisaje de los repositorios institucionales open access en España. *BID: textos universitaris de biblioteconomia i documentació* (20). Recuperado de <http://bid.ub.edu/20meler4.htm>

Monteagudo-Haro, P. & Prieto-Gutierrez, J. J. (2024). Datos abiertos de investigación en repositorios universitarios españoles. *Revista Española de Documentación Científica*, 47(3), e397. <https://doi.org/10.3989/redc.2024.3.1581>

Morcillo López, L. (2016). Los repositorios institucionales en las universidades públicas de España: estado de la cuestión. *Cuadernos de Gestión de Información*, 6, 69–83. Recuperado a partir de <https://revistas.um.es/gesinfo/article/view/264121>

Oficina de proyecto Hércules. (s.f.). *Hércules*. Universidad de Murcia. <https://www.um.es/web/hercules/>

Open Archives Initiative. (s.f.). *The Open Archives Initiative Protocol for Metadata Harvesting*. Recuperado el 21 de marzo de 2025, de <https://www.openarchives.org/OAI/openarchivesprotocol.html>



Reipo, R., Orduña-Malea, E., & Aguaded, I. (2019). Revistas científicas editadas por universidades en Web of Science: características y contribución a la marca universidad. *El profesional de la información*, 28(4). <https://doi.org/10.3145/epi.2019.jul.05>

Segarra, J., & Ortiz, J. (s.f.). *Framework para el soporte a la metodología de Linked Open Data y su aplicación sobre diversos casos de uso*. Universidad de Cuenca. <https://ucuenca.github.io/lodplatform>

Servicio de Publicaciones, Universidad de Murcia. (2013). *Revistas Científicas de la Universidad de Murcia*. <https://revistas.um.es>

Sumba, F., Ortiz, J., Segarra, J., & Saquicela, V. (2017). Integración de fuentes de datos bibliográficas utilizando tecnologías de Linked Data - Caso de uso: Biblioteca de la Universidad de Cuenca. Maskana. <https://publicaciones.ucuenca.edu.ec/ojs/index.php/maskana/article/download/1462/1136>

Universidad de Cuenca. (s.f.). *LOD Platform*. Recuperado el 19 de marzo de 2025 de <https://ucuenca.github.io/lodplatform>

Universidad de La Rioja. (2023). *Crue, la Universidad de Murcia, la Universidad de La Rioja y Dialnet firman un convenio para integrar la gestión de la investigación universitaria en un único portal en español*. Universidad de La Rioja. Recuperado el 25 de febrero de 2025 de <https://www.unirioja.es/acuerdo-crue-umu-ur-hercules-dialnet/>

Xavier, A., & Hernández, F. (2020). OAI-PMH y Linked Open Data en el contexto de Hispana y Europeana: algunas reflexiones históricas. *JLIS.it* 11(1), 1-16. <https://doi.org/10.4403/jlis.it-12573>



## 7. Anexos

A continuación se presentan los anexos correspondientes a este proyecto. Se han incluido materiales complementarios para documentar diversos recursos técnicos y detalles ampliados de algunos resultados.

### Anexo I. Fragmento de Código del script de extracción de OAI-PMH a XML

```
# Archivos de entrada y salida
CSV_FILE = os.path.join(TABLAS, 'revistas.csv')
TXT_FILE = os.path.join(LOGS, 'extraccion_xml_articulos_oai_dc.txt')
XML_FILE = os.path.join(XML, 'articulos_oai_dc.xml')

def extract_raw_oai_pmh_to_xml(base_url, xml_file):
    """
        Extrae todos los registros OAI-PMH en formato oai_dc de una URL base y los guarda
        directamente en el archivo XML,
        excluyendo los registros eliminados.
    """
    url = f"{base_url}/oai?verb=ListRecords&metadataPrefix=oai_dc"
    while True:
        for record in records:
            # Verificar si el registro está marcado como eliminado
            header = record.find('.//{http://www.openarchives.org/OAI/2.0/}header')
            if header is not None and header.get('status') == 'deleted':
                deleted_count += 1
                continue # Saltar este registro
            record_count += 1

            # Buscar el resumptionToken para continuar con la paginación
            resumption_token =
                root.find('.//{http://www.openarchives.org/OAI/2.0/}resumptionToken')
            if resumption_token is None or not resumption_token.text:
                break

            url = f"{base_url}/oai?verb=ListRecords&resumptionToken={resumption_token.text}"
            time.sleep(1) # Pausa para no saturar el servidor
    return record_count, error_count, deleted_count, filtered_count

# Crear el archivo XML con la cabecera inicial
with open(XML_FILE, 'w', encoding='utf-8') as file:
    file.write('<?xml version="1.0" encoding="UTF-8"?>\n')
    file.write('<OAI-PMH>\n')

try:
    with open(csv_file, 'r', encoding='utf-8') as csvfile:
        reader = csv.DictReader(csvfile)
except FileNotFoundError:
    print(f"Error: El archivo {csv_file} no se encuentra.")
    return

# Cerrar la etiqueta raíz del archivo XML al final del proceso
with open(XML_FILE, 'a', encoding='utf-8') as file:
    file.write('</OAI-PMH>\n')
```



## Anexo II. Fragmento del Perfil de aplicación en formato TTL

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix ontorevistas: <http://gicd.inf.um.es/wd/ontorevistas/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix bibo: <http://purl.org/ontology/bibo/> .
@prefix schema: <http://schema.org/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://gicd.inf.um.es/wd/ontorevistas/> a owl:Ontology;
  dc:title "Ontología de Revistas Científicas"@es;
  dc:description "Ontología para representar información sobre revistas científicas, números, artículos y autores."@es;
  owl:imports <http://www.w3.org/2004/02/skos/core>, <http://purl.org/ontology/bibo/>,
    <http://schema.org/>, <http://purl.org/dc/terms/>, <http://xmlns.com/foaf/0.1/>;
  dc:creator "Pedro Otálora";
  dc:created "2025-02-15";
  dc:license "http://creativecommons.org/licenses/by/4.0/" .

ontorevistas:Revista a owl:Class;
  rdfs:subClassOf bibo:Journal;
  rdfs:label "Revista"@es;
  rdfs:comment "Revistas científicas de la Universidad de Murcia"@es;
  owl:equivalentClass schema:Periodical .

ontorevistas:Articulo a owl:Class;
  rdfs:subClassOf bibo:AcademicArticle, bibo:JournalArticle;
  rdfs:label "Articulo"@es;
  owl:equivalentClass schema:ScholarlyArticle .

ontorevistas:Autor a owl:Class;
  rdfs:subClassOf foaf:Person;
  rdfs:label "Autor"@es;
  owl:equivalentClass schema:Person .

ontorevistas:PalabraClave a owl:Class;
  rdfs:subClassOf skos:Concept, bibo:DocumentPart;
  rdfs:label "Palabra clave"@es .

ontorevistas:GrupoTema a owl:Class;
  rdfs:subClassOf skos:Collection;
  rdfs:label "Grupo Temático"@es;
  rdfs:comment "Colección de conceptos que representan un área temática para clasificar artículos científicos"@es .

ontorevistas:Numero a owl:Class;
  rdfs:subClassOf bibo:Issue;
  rdfs:label "Número"@es .

ontorevistas:tieneNumero a owl:ObjectProperty;
  rdfs:label "tiene número"@es;
  rdfs:range ontorevistas:Numero;
  rdfs:domain ontorevistas:Revista;
  owl:equivalentProperty bibo:issue;
  owl:inverseOf ontorevistas:esParteDeRevista .

ontorevistas:esParteDeRevista a owl:ObjectProperty;
  rdfs:label "es parte de revista"@es;
  rdfs:range ontorevistas:Revista;
  rdfs:domain ontorevistas:Numero;
  owl:equivalentProperty dcterms:isPartOf;
  owl:inverseOf ontorevistas:tieneNumero .
```



### Anexo III. Ejemplo de informes del proceso de transformación

Informe	extraccion_xml_articulos_oai_dc.txt
[...] Resumen final: Total de revistas procesadas: 43 Total de registros extraídos: 24650 Total de registros con errores: 0 Total de registros eliminados omitidos: 296 Total de registros filtrados por 'revisores': 40 Tiempo total de ejecución: 0:23:07.206866	

Informe	creacion_tabla_numeros.txt															
[...] Estadísticas: <ul style="list-style-type: none"><li>Total de combinaciones únicas: 1596</li><li>Artículos sin volumen asignado: 0</li><li>Campos vacíos en Revista_ID: 0</li></ul> Primeras 5 entradas: <table><thead><tr><th>Revista_ID</th><th>nombre_volumen</th><th>count_articulos</th></tr></thead><tbody><tr><td>analesbio</td><td>Anales de Biología; Núm. 1 (1984): Sección especial</td><td>31</td></tr><tr><td>analesbio</td><td>Anales de Biología; Núm. 10 (1986): Sección Biología general</td><td>7</td></tr><tr><td>analesbio</td><td>Anales de Biología; Núm. 11 (1987): Sección Biología animal</td><td>11</td></tr><tr><td>analesbio</td><td>Anales de Biología; Núm. 12 (1987): Sección Biología ambiental</td><td>10</td></tr></tbody></table> Tiempo de ejecución: 0:00:00.100459		Revista_ID	nombre_volumen	count_articulos	analesbio	Anales de Biología; Núm. 1 (1984): Sección especial	31	analesbio	Anales de Biología; Núm. 10 (1986): Sección Biología general	7	analesbio	Anales de Biología; Núm. 11 (1987): Sección Biología animal	11	analesbio	Anales de Biología; Núm. 12 (1987): Sección Biología ambiental	10
Revista_ID	nombre_volumen	count_articulos														
analesbio	Anales de Biología; Núm. 1 (1984): Sección especial	31														
analesbio	Anales de Biología; Núm. 10 (1986): Sección Biología general	7														
analesbio	Anales de Biología; Núm. 11 (1987): Sección Biología animal	11														
analesbio	Anales de Biología; Núm. 12 (1987): Sección Biología ambiental	10														

Informe	creacion_tabla_corpus.txt
[...] Resumen de extracción, limpieza y lematización del corpus: ----- Total de registros procesados: 24650 Total de caracteres originales: 54799118 Total de etiquetas HTML eliminadas: 14 Total de saltos de línea eliminados: 9966 Total de palabras antes del procesamiento: 3041430 Total de palabras eliminadas: 1763863 Total de palabras después del procesamiento: 1277567 Reducción de palabras: 57.99% Longitud promedio de los términos finales (en palabras): 52.05 Término más corto (en palabras): "" (0 palabras) Total de stopwords utilizadas: 808 Tiempo total de ejecución: 0:15:57.494754	

Informe	autores_normalizados.txt
Autores originales: 35109 Autores finales: 33213 Grupos fusionados: 1722  Grupo: achury, diana marcela ↳ Variante principal: Achury, Diana Marcela  --- Variantes detectadas (2):        • Achury, Diana Marcela        • Achury, Diana Marcela  --- Artículos consolidados: 3  --- Apariciones totales: 3  [...]	



## Anexo IV. Archivos CSV intermedios creados en el proceso de limpieza

En este apartado se recoge una muestra de algunos de los archivos CSV intermedios generados en los procesos de limpieza y normalización de datos.

**Archivo revistas.csv.** Resultado de la extracción de datos de revistas desde el portal con *web scraping*

Revista_ID	Nombre	URL	Imagen	DOI	ISSN-E	ISSN-Impreso
analesbio	Anales de Biología	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://revistas.um.es/public/j">https://revistas.um.es/public/j</a>	<a href="https://doi.org/10.6018/">https://doi.org/10.6018/</a>	1989-2128	1138-3399
analesderecho	Anales de Derecho	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://revistas.um.es/public/j">https://revistas.um.es/public/j</a>	<a href="https://doi.org/10.6018/">https://doi.org/10.6018/</a>	1989-5992	0210-539X
analedoc	Anales de Documentación	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://revistas.um.es/public/j">https://revistas.um.es/public/j</a>	<a href="https://doi.org/10.6018/">https://doi.org/10.6018/</a>	1697-7904	1575-2437
analesff	Anales de Filología Francesa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://revistas.um.es/public/j">https://revistas.um.es/public/j</a>	<a href="https://doi.org/10.6018/">https://doi.org/10.6018/</a>	1989-4678	0213-2958
analesps	Anales de Psicología / Annals of Psychology	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://revistas.um.es/public/j">https://revistas.um.es/public/j</a>	<a href="https://doi.org/10.6018/">https://doi.org/10.6018/</a>	1695-2294	
analesvet	Anales de Veterinaria de Murcia	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://revistas.um.es/public/j">https://revistas.um.es/public/j</a>	<a href="https://doi.org/10.6018/">https://doi.org/10.6018/</a>	1989-1784	0213-5434
ayc	Antigüedad y Cristianismo	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://revistas.um.es/public/j">https://revistas.um.es/public/j</a>	<a href="https://doi.org/10.6018/">https://doi.org/10.6018/</a>	1989-6182	0214-7165
areas	Áreas. Revista Internacional de Ciencias Sociales	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://revistas.um.es/public/j">https://revistas.um.es/public/j</a>	<a href="https://doi.org/10.6018/">https://doi.org/10.6018/</a>	1989-6190	0211-6707

**Archivo artículos.csv.** Resultado de la extracción y limpieza de los artículos desde el XML

Revista	Revista_ID	url	ns0_ide title	identifier	doi	creator	subject	description	publisher	date	source	language	relation
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	APROXIMACIÓN AL ESTUDIO DE LA VEGETACIÓN EN EL PARQUE NATURAL DE LA SIERRA DE CORDOBA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Galán de Mera, J. M.	vegetación	En este trabajo se ha estudiado...	Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	GERMINACIÓN Y DISTRIBUCIÓN DE SEEDS EN LA MONTAÑA DE LOS BOSQUES EN EL PARQUE NATURAL DE LA SIERRA DE CORDOBA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Escribá, M.	endemismos	Se ha estudiado...	Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	EFFECT OF FIRE ON THE REGENERATION OF A TRIPLEX HALIMIUS COMMUNITIES IN THE SIERRA DE CORDOBA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Nedjimi, B.	Atriplex halimus	Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	ANÁLISIS DE LOS CAMBIOS EN LA DIVERSIDAD DE AVES EN LOS PARQUES NACIONALES DE CORDOBA Y GAVIRA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Gavira, Os	corredor ecológico	Los Parques Nacionales de Gavira y...	Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	ANÁLISIS DE LOS CAMBIOS EN LA DIVERSIDAD DE AVES EN LOS PARQUES NACIONALES DE CORDOBA Y GAVIRA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Carrión García, G.	palinología	El yacimiento arqueológico...	Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	CAMBIOS EN LA DIVERSIDAD DE AVES EN LOS PARQUES NACIONALES DE CORDOBA Y GAVIRA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Yáñez Carrasco, M.	palinología	Este trabajo se ha...	Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	RELACIONES ENTRE LA DIVERSIDAD DE PECES MARINOS Y LA RELACIÓN ENTRE LOS ECOLOGOS	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Yaqoob Al	peces marinos	Relación entre la diversidad...	Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	AN APPROXIMATION TO THE PREDICTION OF THE DIVERSITY OF FISHES IN THE COASTAL AREA OF THE MEDITERRANEAN SEA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Tío, Rober	ochlerotatus	Predicción de la diversidad...	Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	ESTUDIO DE LA DIVERSIDAD DE MOLCILLOS EN EL PARQUE NATURAL DE LA SIERRA DE CORDOBA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Morcillo Alarcón, A.	entomología	Se ha estudiado...	Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	NOVEDADES EN LA DIVERSIDAD DE AVES EN LOS PARQUES NACIONALES DE CORDOBA Y GAVIRA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Cano, María J.   Guerra, Juan			Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	NUEVAS REFERENCIAS SOBRE LA DIVERSIDAD DE AVES EN LOS PARQUES NACIONALES DE CORDOBA Y GAVIRA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Sánchez Gómez, Pedro   Jiménez, Juan			Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	EL PROFESOR DE BIOLOGIA EN LOS COLEGIOS DE LA PROVINCIA DE CORDOBA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Argüelles, Juan Carlos		A comienzos del...	Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	NOVEDADES EN LA DIVERSIDAD DE AVES EN LOS PARQUES NACIONALES DE CORDOBA Y GAVIRA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Anales de Biología, Revista			Facultad de Biología	2006-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	EFFECT OF SALT POLLUTION ON THE DIVERSITY OF FISHES IN THE MEDITERRANEAN SEA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	Damjibhai	salinisation	Effects of salinity...	Facultad de Biología	2005-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>
Anales de Biología	analesbio	<a href="https://oai:revistas.um.es/">https://oai:revistas.um.es/</a>	FENOLOGÍA EN LOS PARQUES NACIONALES DE CORDOBA Y GAVIRA	<a href="https://revistas.um.es/">https://revistas.um.es/</a>	<a href="https://doi.org/10.6018/0000-0000-0000-0000">https://doi.org/10.6018/0000-0000-0000-0000</a>	García Peña, A.	Acrocephalus	Se muestra la...	Facultad de Biología	2005-12	Anales de Biología	spa	<a href="https://revistas.um.es/">https://revistas.um.es/</a>

**Archivo números.csv.** Resultado de procesar el metadato <dc:source> para identificar el número/volumen al que pertenece el artículo.

Revista_ID	nombre_volumen	count_articulos
analesbio	Anales de Biología; Núm. 1 (1984): Sección especial	31
analesbio	Anales de Biología; Núm. 10 (1986): Sección Biología general	7
analesbio	Anales de Biología; Núm. 11 (1987): Sección Biología animal	11
analesbio	Anales de Biología; Núm. 12 (1987): Sección Biología ambiental	10
analesbio	Anales de Biología; Núm. 13 (1987): Sección Biología vegetal	11
analesbio	Anales de Biología; Núm. 14 (1987): Sección Biología general	8
analesbio	Anales de Biología; Núm. 15 (1988-89)	24
analesbio	Anales de Biología; Núm. 16 (1990)	21



## Transformación de metadatos OAI-PMH a conjuntos de datos semánticos RDF en revistas científicas y repositorios institucionales

**Archivo autores\_normalizados.csv.** Resultado de la extracción y normalización de autores.

author	ns0_identifiers	0
"Enfermería Global", Revista	oai:revistas.um.es:article/115931	1
(Editum), Ediciones de la Universidad de Murcia	oai:revistas.um.es:article/345951	1
(GERECS), Grupo de editores de revistas españolas	oai:revistas.um.es:article/315191	1
Cartaphilus	oai:revistas.um.es:article/92   oai:revistas.um.es:article/94	2
Consejo de Redacción	oai:revistas.um.es:article/231591   oai:revistas.um.es:article/198971   oai:revistas.um.es:article/216561	6
RIE	oai:revistas.um.es:article/98911	1
A. Geiger, Marshall	oai:revistas.um.es:article/388671	1

**Archivo keywords\_unesco.csv.** Resultado de la reconciliación de palabras clave con el tesoro de la Unesco.

keyword	unesco_URI	idioma
abastecimiento de agua	<a href="http://vocabularies.unesco.org/thesaurus/concept985">http://vocabularies.unesco.org/thesaurus/concept985</a>	es
ability	<a href="http://vocabularies.unesco.org/thesaurus/concept993">http://vocabularies.unesco.org/thesaurus/concept993</a>	en
aborto	<a href="http://vocabularies.unesco.org/thesaurus/concept1010">http://vocabularies.unesco.org/thesaurus/concept1010</a>	es
abstracting	<a href="http://vocabularies.unesco.org/thesaurus/concept1022">http://vocabularies.unesco.org/thesaurus/concept1022</a>	en
abuso sexual	<a href="http://vocabularies.unesco.org/thesaurus/concept1035">http://vocabularies.unesco.org/thesaurus/concept1035</a>	es
academic achievement	<a href="http://vocabularies.unesco.org/thesaurus/concept87">http://vocabularies.unesco.org/thesaurus/concept87</a>	en
acceso a la educación	<a href="http://vocabularies.unesco.org/thesaurus/concept1124">http://vocabularies.unesco.org/thesaurus/concept1124</a>	es

**Archivo tesauro\_unesco\_corpus.csv.** Resultado de aplicar técnicas de PLN a la cadena de texto formada por todos los términos de cada microtesauro.

Grupo	Grupo_	Grupo_URI	MicroTesauro	MicroT	MicroTesauro_L	Términos
Educación	1	<a href="http://vocabulari">http://vocabulari</a>	Ciencias de la educación y ambiente edu	1.05	<a href="http://vocabulari">http://vocabulari</a>	psicología educación psicología educativo psicopedagog
Educación	1	<a href="http://vocabulari">http://vocabulari</a>	Política educacional	1.10	<a href="http://vocabulari">http://vocabulari</a>	derecho educación oportunidad educacional democratiza
Educación	1	<a href="http://vocabulari">http://vocabulari</a>	Planificación de la educación	1.15	<a href="http://vocabulari">http://vocabulari</a>	asistencia escolar frecuentación colegio presenciar colegi
Educación	1	<a href="http://vocabulari">http://vocabulari</a>	Administración de la educación	1.20	<a href="http://vocabulari">http://vocabulari</a>	libertad enseñanza derechos civil situación docente situaci
Educación	1	<a href="http://vocabulari">http://vocabulari</a>	Gestión de la educación	1.25	<a href="http://vocabulari">http://vocabulari</a>	agrupamiento aptitud división función aptitud grupo nivel
Educación	1	<a href="http://vocabulari">http://vocabulari</a>	Sistemas y niveles de enseñanza	1.30	<a href="http://vocabulari">http://vocabulari</a>	servicio educativo itinerante aula móvil educacional itiner
Educación	1	<a href="http://vocabulari">http://vocabulari</a>	Establecimientos de enseñanza	1.35	<a href="http://vocabulari">http://vocabulari</a>	escuela pequeño colegio aula tamaño escuela universitari
Educación	1	<a href="http://vocabulari">http://vocabulari</a>	Planes de estudios	1.40	<a href="http://vocabulari">http://vocabulari</a>	curso acelerado actividad programa actividad actividad tí

**Archivo clusters.csv.** Resultado de comparar el corpus de artículos con el del tesauro para asignar una categoría al artículo. En la columna “Corpus” se encuentra el texto procesado con técnicas de PLN. En la columna “Términos\_Clave” se pueden ver los términos clave que coinciden con el corpus del tesauro asignado.

ns_identifier	Corpus	Categoría1	cluster	Terminos_Clave
<a href="#">oai:revistas.um.es:ar</a>	aproximación esquema vegetación región sur trabajo	Africa	7.05	sur
<a href="#">oai:revistas.um.es:ar</a>	germinación endemismo provincia estudiar germinaci	Ciencias naturales	2.75	vegetal, especie
<a href="#">oai:revistas.um.es:ar</a>	effect soluble effect soluble conditions soluble result	Sin clasificación	0	
<a href="#">oai:revistas.um.es:ar</a>	análisis corredor parque natural sierra sierra nieve par	Ciencias naturales	2.75	natural, flora, parque, distribución
<a href="#">oai:revistas.um.es:ar</a>	análisis polínico yacimiento arqueológico yacimiento	Historia	3.25	arqueológico, yacimiento, contemporáneo
<a href="#">oai:revistas.um.es:ar</a>	cambio vegetación franja marisma holoceno reciente	Geografía y oceanografía	2.4	marino, estuario, vegetación, franja, e
<a href="#">oai:revistas.um.es:ar</a>	relationship structur iraq relación tamaño corporal est	Geografía y oceanografía	2.4	marino, zona, forma
<a href="#">oai:revistas.um.es:ar</a>	predicción emergencia estival basado relación acumul	Planes de estudios	1.4	programa, estudio, verano, integrado



## Anexo V. Fragmento de Dataset final en formato TTL

```
datarevistas:revista_analesdoc a ontorevistas:Revista ;
    rdfs:label "Anales de Documentación"^^xsd:string ;
    ontorevistas:revistaDOI <https://doi.org/10.6018/analesdoc> ;
    ontorevistas:revistaISSN "1575-2437"^^xsd:string ;
    ontorevistas:revistaISSNE "1697-7904"^^xsd:string ;
    ontorevistas:revistaImagen
<https://revistas.um.es/public/journals/18/journalThumbnail_es_ES.jpg> ;
    ontorevistas:revistaURL <https://revistas.um.es/analesdoc> ;
    ontorevistas:tieneNumero datarevistas:numero_analesdoc_0093,
        datarevistas:numero_analesdoc_0094,
        datarevistas:numero_analesdoc_0095,
        [...]

datarevistas:numero_analesdoc_0094 a ontorevistas:Numero ;
    rdfs:label "Anales de Documentación; Vol. 10 (2007)"^^xsd:string ;
    ontorevistas:esParteDeRevista datarevistas:revista_analesdoc ;
    ontorevistas:numeroVolumen "Anales de Documentación; Vol. 10 (2007)"^^xsd:string ;
    ontorevistas:tieneArticulo datarevistas:articulo_1082,
        datarevistas:articulo_1092,
        datarevistas:articulo_1102,
        datarevistas:articulo_1111,
        [...]

datarevistas:articulo_1092 a ontorevistas:Articulo ;
    rdfs:label "La voz bibliografía en la enciclopedia universal ilustrada hispanoamericana
de la editorial espasa."^^xsd:string ;
    ontorevistas:articuloEditor "Ediciones de la Universidad de Murcia (Editum)"^^xsd:string ;
    ontorevistas:articuloFechaPublicacion "2008-02-12"^^xsd:date ;
    ontorevistas:articuloIdioma "spa"^^xsd:string ;
    ontorevistas:articuloOAI <oai:revistas.um.es:article/1092> ;
    ontorevistas:articuloRecursoURI
<https://revistas.um.es/analesdoc/article/view/1092/1142> ;
    ontorevistas:articuloResumen "La enciclopedia Espasa constituyó [...] "^^xsd:string ;
    ontorevistas:articuloURL <https://revistas.um.es/analesdoc/article/view/1092> ;
    ontorevistas:esParteDeNumero datarevistas:numero_analesdoc_0094 ;
    ontorevistas:perteneceAGrupoTema datarevistas:tema_5.3 ;
    ontorevistas:tieneAutor datarevistas:autor_AU-UM-01009110 ;
    ontorevistas:tienePalabraClave datarevistas:palabraclave_KW-UM-01010826,
        datarevistas:palabraclave_KW-UM-01012585,
        datarevistas:palabraclave_KW-UM-01013777,
        [...]

datarevistas:autor_AU-UM-01009110 a ontorevistas:Autor ;
    rdfs:label "Fernández Fuentes, Belén"^^xsd:string ;
    ontorevistas:autorID "AU-UM-01009110"^^xsd:string ;
    ontorevistas:esAutorDe datarevistas:articulo_1092 .

datarevistas:palabraclave_KW-UM-01013777 a ontorevistas:PalabraClave ;
    rdfs:label "españa"^^xsd:string ;
    ontorevistas:esPalabraClaveDe datarevistas:articulo_1092,
        datarevistas:articulo_109561,
        datarevistas:articulo_109571,
        [...]

datarevistas:tema_5.3 a ontorevistas:GrupoTema ;
    rdfs:label "Fuentes de información"^^xsd:string ;
    ontorevistas:esGrupoTemaDe datarevistas:articulo_1092,
        datarevistas:articulo_109921,
        datarevistas:articulo_110031,
        [...]
```



## Anexo VI. Fragmento de Código PHP de la aplicación

Ejemplo de código PHP de aplicación

```
<?php
// Incluimos los archivos necesarios
include 'sparqlquerydispatcher.php';
include 'sparql_prefijos.php';
include 'menu_principal.php';
include 'articulos-buscador_sparql.php';

// Parámetros de búsqueda
$filtros = isset($_GET['filtros']) ? $_GET['filtros'] : [];
$resultadosPorPagina = isset($_GET['resultadosPorPagina']) ?
intval($_GET['resultadosPorPagina']) : 10;
$paginaActual = isset($_GET['pagina']) ? intval($_GET['pagina']) : 1;

// Definimos el endpoint SPARQL
include 'endpoint_url.php';
$queryDispatcher = new SPARQLQueryDispatcher($endpointUrl);

// Consulta para contar el total de artículos según los filtros
$sparqlQueryTotal = contarTotalArticulos($filtros);
$queryResultsTotal = $queryDispatcher->query($sparqlQueryTotal);
<?php
// Serializar los filtros para incluirlos en las URLs
$filtrosQuery = !empty($_GET['filtros']) ? http_build_query(['filtros' =>
$_GET['filtros']] : '';

// Número máximo de botones de página a mostrar
$maxBotones = 5;

// Calcular el inicio y el fin del rango de páginas a mostrar
$inicio = max(1, $paginaActual - floor($maxBotones / 2));
$fin = min($totalPaginas, $inicio + $maxBotones - 1);
$inicio = max(1, $fin - $maxBotones + 1);
```

Ejemplo de código para la generación de consultas SPARQL dinámicas

```
<?php
include 'sparql_prefijos.php';

function obtenerArticulos($offset, $limit, $filtros) {
    global $sparqlPrefijos;

    // Generar los filtros dinámicamente en base a los criterios seleccionados
    $sparqlFiltros = '';
    foreach ($filtros as $filtro) {
        if (!empty($filtro['valor'])) {
            // Asegurarse de que el valor sea una cadena
            $valor = is_array($filtro['valor']) ? implode(", ", $filtro['valor']) :
$filtro['valor'];

            switch ($filtro['entidad']) {
                case 'Articulo':
                    $sparqlFiltros .= "FILTER(CONTAINS(LCASE(?ArticuloLabel), LCASE('" .
addslashes($valor) . "')))\n";
                    break;
                case 'Autor':
                    $sparqlFiltros .= "FILTER(EXISTS { ?Articulo ontorevistas:tieneAutor
?Autor . ?Autor rdfs:label ?AutorLabel . FILTER(CONTAINS(LCASE(?AutorLabel), LCASE('" .
addslashes($valor) . "'))) })\n";
                    break;
                case 'PalabraClave':
                    $sparqlFiltros .= "FILTER(EXISTS { ?Articulo
ontorevistas:tienePalabraClave ?PalabraClave . ?PalabraClave rdfs:label ?PalabraClaveLabel .
FILTER(CONTAINS(LCASE(?PalabraClaveLabel), LCASE('" . addslashes($valor) . "'))) })\n";
            }
        }
    }
}
```



## Transformación de metadatos OAI-PMH a conjuntos de datos semánticos RDF en revistas científicas y repositorios institucionales

```
        break;
    case 'Tema':
        $sparqlFiltros .= "FILTER(EXISTS { ?Articulo
ontorevistas:perteneceAGrupoTema ?Tema . ?Tema rdfs:label ?TemaLabel .
FILTER(CONTAINS(LCASE(?TemaLabel), LCASE('". addslashes($valor) . "'))) })\n";
        break;
    case 'Revista':
        $sparqlFiltros .= "FILTER(EXISTS { ?Articulo
ontorevistas:esParteDeNumero/ontorevistas:esParteDeRevista ?Revista . ?Revista rdfs:label
?RevistaLabel . FILTER(CONTAINS(LCASE(?RevistaLabel), LCASE('". addslashes($valor) . "'))) })\n";
        break;
    }
}

// Construir la consulta SPARQL
return <<<SPARQL
$sparqlPrefijos
SELECT DISTINCT ?Articulo ?ArticuloLabel ?RevistaLabel ?NumeroVolumen
  (GROUP_CONCAT(DISTINCT ?AutorLabel; SEPARATOR=", ") AS ?Autores)
  ?FechaPublicacion
WHERE {
  # Artículo y sus propiedades principales
  ?Articulo a ontorevistas:Articulo .

  # Propiedades opcionales
  OPTIONAL { ?Articulo rdfs:label ?ArticuloLabel . }
  OPTIONAL {
    ?Articulo ontorevistas:esParteDeNumero ?Numero .
    ?Numero ontorevistas:esParteDeRevista ?Revista .
    ?Revista rdfs:label ?RevistaLabel .
  }
  OPTIONAL {
    ?Articulo ontorevistas:esParteDeNumero ?Numero .
    ?Numero ontorevistas:numeroVolumen ?NumeroVolumen .
  }
  OPTIONAL { ?Articulo ontorevistas:articuloFechaPublicacion ?FechaPublicacion . }

  # Relación con autores (opcional)
  OPTIONAL {
    ?Articulo ontorevistas:tieneAutor/rdfs:label ?AutorLabel .
  }

  # Aplicar los filtros dinámicos
  $sparqlFiltros
}
GROUP BY ?Articulo ?ArticuloLabel ?RevistaLabel ?NumeroVolumen ?FechaPublicacion
ORDER BY LCASE(?ArticuloLabel)
OFFSET $offset
LIMIT $limit
SPARQL;
}

// Construir la consulta SPARQL para contar los artículos
return <<<SPARQL
$sparqlPrefijos
SELECT (COUNT(DISTINCT ?Articulo) AS ?total)
WHERE {
  # Artículo y sus propiedades principales
  ?Articulo a ontorevistas:Articulo ;
             rdfs:label ?ArticuloLabel .

  # Aplicar los filtros dinámicos
  $sparqlFiltros
}
SPARQL;
}
```



## Anexo VII. Glosario de términos y acrónimos

**Apache Jena Fuseki:** Servidor de código abierto que permite almacenar, gestionar y consultar datos RDF mediante el lenguaje SPARQL, proporcionando un punto de acceso web para la publicación y consulta de datos semánticos.

**BIBO (Bibliographic Ontology):** Ontología en RDF para describir recursos bibliográficos como libros, artículos y revistas.

**Corpus:** En el contexto del procesamiento del lenguaje natural y la inteligencia artificial, un corpus se utiliza para analizar, entrenar y evaluar modelos automáticos, ya que permite identificar patrones lingüísticos relevantes para la investigación.

**CKAN (Comprehensive Knowledge Archive Network):** Plataforma de código abierto diseñada para la creación y gestión de portales de datos abiertos. Permite a organizaciones públicas y privadas publicar, compartir y reutilizar conjuntos de datos de manera estructurada y accesible.

**CSV (Comma-Separated Values):** Formato de archivo de texto plano para almacenar datos tabulares, donde los valores están separados por comas.

**DC (Dublin Core):** Estándar de metadatos para describir recursos digitales.

**DCAT (Data Catalog Vocabulary):** Vocabulario RDF para describir catálogos de datos.

**DCTERMS (Dublin Core Terms):** Extensión del estándar Dublin Core que amplía y detalla los elementos de metadatos.

**DIM (DSpace Intermediate Metadata):** Esquema de metadatos intermedio utilizado en la plataforma DSpace.

**DOI (Digital Object Identifier):** Identificador único y permanente para objetos digitales, como artículos científicos.



**DSpace:** Plataforma de software libre para la gestión y preservación de repositorios institucionales.

**EDM (Europeana Data Model):** Modelo de datos utilizado por Europeana para estructurar y describir metadatos de objetos digitales culturales, facilitando la interoperabilidad entre instituciones.

**Europeana:** Biblioteca digital europea que centraliza y difunde el patrimonio cultural digitalizado de instituciones de toda Europa.

**Expresión regular:** Secuencia de caracteres que se utiliza como máscara para localizar y reemplazar patrones dentro de cadenas de texto.

**FOAF (Friend of a Friend):** Vocabulario RDF para describir personas, sus actividades y sus relaciones.

**GitHub:** Plataforma de desarrollo colaborativo basada en control de versiones para alojar y gestionar proyectos de software.

**Google Colab (Google Colaboratory):** Plataforma gratuita de Google que permite escribir y ejecutar código Python en un entorno de cuadernos interactivos basado en la nube.

**Hispana:** Portal y agregador nacional que reúne y da acceso al patrimonio digital de archivos, bibliotecas y museos españoles. Actúa como recolector y repositorio OAI-PMH, facilitando su integración en Europeana.

**ISSN (International Standard Serial Number):** Código internacional que identifica de forma única publicaciones seriadas.

**ISSNE (ISSN electrónico):** Versión electrónica del ISSN para publicaciones digitales.

**LEM (Lista de Encabezamientos de Materia):** Tesauro para la catalogación temática en bibliotecas públicas españolas.



**Linked Open Data (LOD):** Datos abiertos y enlazados mediante tecnologías semánticas para facilitar su interconexión y reutilización.

**LOV (Linked Open Vocabularies):** Repositorio de vocabularios abiertos para datos enlazados. Más información en <http://lov.linkeddata.es/>.

**MARC (Machine-Readable Cataloging):** Formato estándar para la representación e intercambio de información bibliográfica.

**MARCXML:** Variante en XML del formato MARC.

**METS (Metadata Encoding and Transmission Standard):** Estándar para la codificación y transmisión de metadatos en bibliotecas digitales.

**MODS (Metadata Object Description Schema):** Esquema XML para la descripción bibliográfica de recursos.

**MySQL:** Sistema de gestión de bases de datos relacionales de código abierto que permite almacenar, organizar y consultar datos en tablas mediante el lenguaje SQL (Structured Query Language).

**NLTK (Natural Language Toolkit):** Librería de Python para el procesamiento y análisis de lenguaje natural.

**OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting):** Protocolo para la recolección automatizada de metadatos en repositorios digitales. Más información en <https://www.openarchives.org/pmh/>.

**OJS (Open Journal Systems):** Software de código abierto para la gestión editorial de revistas científicas.

**OpenRefine:** Herramienta de código abierto para la limpieza, transformación y reconciliación de datos provenientes de diversas fuentes y formatos. Más información en <https://openrefine.org/>.



**ORCID (Open Researcher and Contributor ID):** Identificador único para investigadores y autores científicos.

**OWL (Web Ontology Language):** Lenguaje para la definición de ontologías en la web semántica.

**Pentaho:** Plataforma de Business Intelligence (BI) de código abierto que proporciona herramientas integradas para la extracción, transformación y carga de datos (ETL).

**PHP (Hypertext Preprocessor):** Lenguaje de programación de propósito general ampliamente usado en desarrollo web, que se ejecuta en el lado del servidor.

**PLN (Procesamiento de Lenguaje Natural):** Rama de la inteligencia artificial dedicada al tratamiento automático del lenguaje humano.

**Python:** Lenguaje de programación de alto nivel, interpretado y de código abierto. Ampliamente utilizado en desarrollo web, automatización, ciencia de datos, inteligencia artificial y muchas otras áreas.

**QDC (Qualified Dublin Core):** Perfil cualificado del estándar Dublin Core con elementos adicionales.

**RapidFuzz:** Librería de Python para la comparación eficiente y rápida de cadenas de texto mediante emparejamiento difuso.

**RDF (Resource Description Framework):** Modelo estándar para la representación de datos enlazados en la web semántica.

**RDFLib:** Librería de Python para trabajar con datos RDF.

**RDFS (RDF Schema):** Extensión de RDF para describir vocabularios y estructuras de datos.

**RECOLECTA:** Infraestructura nacional que agrupa repositorios científicos españoles de acceso abierto, fruto de la colaboración entre la Fundación Española para la Ciencia y



la Tecnología (FECYT) y la Red de Bibliotecas Universitarias (REBIUN). Promueve la interoperabilidad y el acceso abierto, facilitando la visibilidad de la investigación científica.

**RFC1807:** Especificación para la descripción de documentos técnicos y científicos en formato MARC.

**Schema.org:** Colección de esquemas de metadatos para estructurar información en la web.

**SKOS (Simple Knowledge Organization System):** Modelo para la representación de esquemas de conceptos, tesauros y taxonomías.

**spaCy:** Librería de Python para procesamiento avanzado y eficiente de lenguaje natural.

**SPARQL (SPARQL Protocol and RDF Query Language):** Lenguaje de consulta para bases de datos RDF.

**TF-IDF (Term Frequency-Inverse Document Frequency):** Técnica para ponderar la importancia de palabras en un corpus de documentos.

**Tripletas:** Estructura básica de datos en el modelo RDF compuesta por tres elementos: sujeto, predicado y objeto, que representa una relación entre recursos de la web semántica.

**TTL (Turtle RDF Triple Language):** Formato sintáctico compacto para expresar datos RDF.

**UKETD\_DC (UK Electronic Theses and Dissertations Dublin Core):** Perfil de metadatos Dublin Core para tesis doctorales del Reino Unido.

**UNESCO (United Nations Educational, Scientific and Cultural Organization):** Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura.



**URL (Uniform Resource Locator):** Dirección única que localiza un recurso en Internet.

**URI (Uniform Resource Identifier):** Identificador único de un recurso en la web.

**Verbos OAI-PMH:** Conjunto de comandos definidos por el protocolo OAI-PMH para la recuperación de metadatos mediante solicitudes HTTP. Cada verbo corresponde a una función específica, permitiendo recuperar desde información del repositorio a registros individuales. Los verbos son: *Identify*, *ListMetadataFormats*, *ListIdentifiers*, *ListRecords*, *GetRecord* y *ListSets*.

**VocBench:** Herramienta de código abierto para el desarrollo de vocabularios controlados y ontologías en un entorno gráfico.

**Web scraping:** Técnica para la extracción automatizada de datos estructurados de páginas web (DeVito et al., 2020).

**Web semántica:** Extensión de la web tradicional que incorpora metadatos y ontologías para dotar de significado a los datos, permitiendo su procesamiento automático por máquinas y mejorando la interoperabilidad entre sistemas.

**XML (eXtensible Markup Language):** Lenguaje de marcado extensible para estructurar datos.

**XOAI (Extensible OAI):** Extensión del protocolo OAI-PMH utilizada en DSpace.

