Revista PANAMERICANA de COMUNICACIÓN

Vol. 6, n. 1, enero-junio. 2024

Proyecto de web semántica de autoridades en PARES: extracción y análisis inicial

Project of semantic web of PARES authorities: extraction and initial analysis

Manuel Blázquez-Ochando; María-Antonia Ovalle-Perandones

Manuel Blázquez-Ochando
Universidad Complutense de Madrid, España
manublaz@ucm.es
https://orcid.org/0000-0002-4108-7531

María-Antonia Ovalle-Perandones
Universidad Complutense de Madrid, España
maovalle@ucm.es
https://orcid.org/0000-0002-6149-4724

Cómo citar este artículo

Blázquez-Ochando, M., & Ovalle-Perandones, M. A. (2023). Proyecto de web semántica de autoridades en PARES: extracción y análisis inicial. *Revista Panamericana de Comunicación*, *6*(1), 1-13. https://doi.org/10.21555/rpc.v6i1.3121

Recibido: 31 - 03 - 2024
Aceptado: 31 - 05 - 2024
Publicado en línea: 20 - 06 - 2024

Resumen

La investigación se centra en describir los tipos de autoridades del Portal de Archivos Españoles, aportando su cuantificación, y ratio relacional, con el fin de delinear el grafo inicial de este sector en PARES. Para lograrlo se emplean métodos de webscraping que han permitido la compilación de todos los registros de autoridad, para su procesamiento y análisis. Los datos recopilados muestran mayor relevancia de las autoridades personales y familias, seguidas de instituciones y conceptos. Este enfoque refleja la importancia de los individuos y las relaciones familiares en el contexto histórico y archivístico. Además, se destacan las relaciones asociativas entre personas e instituciones, lo que sugiere la complejidad de las interacciones sociales y organizacionales en el pasado. También se comprueba una fuerte interconexión entre lugares y personas, así como entre lugares y otras entidades como instituciones y normas. Esto subraya la importancia de la geolocalización y el contexto geográfico en la comprensión del patrimonio histórico y cultural representado en PARES. Además, se identifica una proporción equitativa entre relaciones familiares, lo que indica una representación rica de la vida social y familiar. Por otro lado, se observa una baja proporción de relaciones asociativas con fuentes de información, lo que sugiere la necesidad de ampliar la documentación y las referencias utilizadas en las fichas descriptivas.

Palabras clave: Web Semántica; PARES; Autoridades; Webscraping; Modelo relacional; Grafo del conocimiento; Interacciones sociales; Patrimonio histórico; Fuentes de información.

Abstract

The research focuses on describing the types of authorities within the Spanish Archives Portal, providing their quantification and relational ratio to outline the initial graph of this sector in PARES. To achieve this, web scraping methods have been employed, allowing for the compilation of all authority records for processing and analysis. The collected data demonstrates the greater relevance of personal authorities and families, followed by institutions and concepts. This approach reflects the importance of individuals and family relationships in the historical and archival context. Additionally, associative relationships between individuals and institutions are highlighted, suggesting the complexity of social and organizational interactions in the past. Furthermore, a strong interconnection between places and individuals is observed, as well as between places and other entities such as institutions and norms. This underscores the importance of geolocation and geographic context in understanding the historical and cultural heritage represented in PARES. Moreover, an equitable proportion of family relationships is identified, indicating a rich representation of social and family life. Conversely, there is a low proportion of associative relationships with information sources, suggesting the need to expand the documentation and references used in descriptive records.

Keywords: Semantic Web; PARES (Spanish Archives Portal); Authorities; Webscraping; Relational model; Knowledge graph; Social interactions; Historical heritage; Information sources.

Introducción

La denominada Web 1.0, estática, hipertextual o de documentos, integra el uso de protocolos como http, FTP, email, *world wide web*, entre otros. Desde la perspectiva de los documentos, se considerará esa web como el entorno tecnológico que facilita el intercambio de información a través de un cliente o navegador y el usuario que accede a ella gracias a lo facilitado por un determinado servidor en el que permanecen almacenados los documentos y la información en ellos registrada. Aunque también existe la Web 2.0 o social (O'Reilly, 2005), centramos la atención en Web semántica como parte de la Web 3.0. En ella toman significado los datos y, sobre ellos, se construye el conocimiento. Es importante considerar que en esta concepción de la web se usa la semántica para crear vocabularios de dominio que fomentan la inferencia de las aplicaciones de software sirviéndose, por ejemplo, de las ontologías. Eso es lo que permite pasar de la web de documentos (HTML, principalmente) a la web de datos, con una intermediaria que es más dinámica y social. La semántica en ese contexto se caracteriza por ser un espacio en el que los datos se expresan en un lenguaje especial, con datos formales y semánticamente interconectados con otros datos.

El modelo de capas definido por Tim Berners-Lee se conoce como *Semantic Web Layer Cake* (Miller, 2001). La primera versión data de 1998, si bien es una representación frecuentemente revisada. En ella, en la primera capa, aparecen los URI y la codificación universal Unicode. Tras ella, la capa que cuenta con las capacidades del metalenguaje XML para definir esquemas de marcado personalizados que permite añadir pequeñas proporciones de semántica a datos ajustados a una estructura. Ascendiendo hasta la tercera capa, se sitúa RDF o *Resource Description Framework* que es el marco para la representación de información en la web. Está basado en un modelo de datos abstracto y se expresa a través de XML mediante una sintaxis denominada RDF/XML. Después, RDFSchema, si RDF es sólo un marco general para la descripción de recursos, es necesario especificar y restringir las descripciones para que se adapten a las necesidades concretas de aplicación de metadatos propios de diversos dominios. Los *schema* RDF permiten la creación de vocabularios específicos para campos de aplicación concretos. Su capacidad expresiva es más amplia con respecto del marco, pero limitada. Por esa razón, en las capas superiores hay que considerar *OWL Web Ontology Language* como el lenguaje para la construcción de ontologías. Así, una ontología se puede definir como un conjunto de especificaciones de conceptos que tienen la capacidad de ser utilizadas por máquinas y así describir estructuras conceptuales complejas. Después, SPARQL o *Simple Protocol And RDF Query Language* que es un lenguaje similar a SQL, pero que utiliza tripletas y expresiones RDF para hacer coincidir parte de la consulta y entonces devolver los resultados pertinentes. Dado que tanto RDFS como OWL están construidos en RDF, SPARQL puede usarse asimismo para consultar ontologías y bases de conocimiento. Hay que considerar que SPARQL no es sólo un lenguaje de consulta, sino que también es un protocolo para acceder a datos ajustados al marco RDF. El resto de las capas y componentes tienen como finalidad soportar otros propósitos, como la inferencia, las declaraciones fiables o fiabilidad.

Los datos expresados con el conjunto de tecnologías integradas en la *Semantic Web Layer Cake* son los denominados datos enlazados (*Linked Data*, LD) que en ocasiones son mencionados sumando el matiz denominado datos abiertos enlazados (*Linked Open Data*, LOD). Los últimos se basan en los datos expresados con las mencionadas tecnologías, pero quedando vinculados a otros conjuntos de datos y así formando grafos del conocimiento (Hogan et al., 2021) derivados de los conjuntos interrelacionados. Uno de esos grafos, entre otros, es la nube o LOD cloud:
https://lod-cloud.net

Los únicos conjuntos de datos integrados en la LOD Cloud y que guardan algún nexo con el contexto de los archivos como instituciones documentales son:

- Archives Hub
  https://lod-cloud.net/dataset/archiveshub-linkeddata
- Archive of the Art Textbooks of Elementary and Public Schools in the Japanese Colonial Period
  https://lod-cloud.net/dataset/ASCDC-AS-NTUE-School-Art-Textbooks

Las instituciones documentales identificadas con el acrónimo LAM (*Libraries, Archives and Museums*, que aglutina a instituciones bibliotecarias, archivísticas y museísticas), en lo referente a web semántica o LD en España se han centrado principalmente en las bibliotecas. Son referentes los datos enlazados bibliotecarios en ese dominio geográfico, los de la Biblioteca Nacional de España datos.bne.es, los de la Biblioteca Virtual Miguel de Cervantes data.cervantesvirtual.com, la Biblioteca Escolar Digital del Centro Internacional de Tecnologías Avanzadas (CITA) o el proyecto no vigente desarrollado en el contexto de la Universidad Pontificia de Salamanca (UPSA), anteriormente disponible en el dominio dataupsa.upsa.es.

Las iniciativas basadas en datos enlazados para archivos son de menor alcance. En España hay que atribuir los avances realizados en esta materia a proyectos en colaboración como los desarrollados por los archivos municipales, entre otros el Archivo del Ayuntamiento de Arganda del Rey o el del Archivo Municipal del Ayuntamiento de Burgos. En otro nivel administrativo también cabe la mención al caso de Documentos y Archivos de Aragón: DARA. En otro ámbito, la Conferencia de Rectores de Universidades Españolas (CRUE, 2017) organizó el Grupo de Trabajo *Linked Open Data* y Archivos Universitarios publicando una Guía LOD para esos archivos, si bien hasta la fecha no ha surgido ninguna iniciativa o no se ha difundido. En Europa, la *OAD ontology* ha permitido desarrollar proyectos LOD en el Archivio storico della Presidenza della Repubblica o Archivi della scienza, entre otros (Guernaccini et al., 2019). La *ArDO ontology* y *RiC-O* han soportado proyectos como el desarrollado en Weimar Republic (Vafaie et al. 2021).

El avance es incipiente y parece que los archivos no se han abierto al potencial que tiene la publicación de datos registrados en los documentos en archivos españoles, frente al avance en otros contextos geográficos diferentes como Reino Unido (Maynard & Greenwood, 2012) o Portugal (Koch et al., 2019). Mientras eso ocurre, todo sugiere que los datos con semántica para los archivos concentran su atención en los datos geográficos (Jacobs et al., 2015). A ellas también se debe sumar el grupo de iniciativas que emanan del Archives Portal Europe Foundation (APEF, n.d.), que trabaja con esquemas como EAD.

El recorrido a realizar para ir desde la interoperabilidad hacia la interoperabilidad semántica en archivos debe situarse en el modelo *Records in Contexts: A Conceptual Model for Archival Description* (RiC-CM) (López Cuadrado & Requejo Zalama, 2021) de manera similar a como la interoperabilidad semántica en bibliotecas se fundamentó en FRBR (Llanes-Padrón & Pastor-Sánchez, 2017).

En el sistema archivístico español, se consolidó hace años como una herramienta fundamental para la recuperación de datos archivísticos el Portal de Archivos Españoles, también conocido como PARES. Con él, se dispone de una plataforma que elaboró y mantiene el Ministerio que tiene asignadas las competencias en materia de Cultura en España. Su fin es dar difusión a toda la información y datos archivísticos del Patrimonio Histórico Documental Español. Esos datos provienen de la red formada por los once archivos que son de titularidad estatal. Estos son:

- Archivos de la Corona de Aragón (ACA);
- Archivo Central del Ministerio de Cultura (ACC);
- Archivo General de la Administración (AGA);
- Archivo General de Indias (AGI);

- Archivo General de Simancas (AGS);
- Archivo Histórico Nacional (AHN);
- Archivo Histórico de la Nobleza (ANOB);
- Archivo de la Real Chancillería de Valladolid (ARCH);
- Centro Documental de la Memoria Histórica (CDMH);
- Centro de Información Documental de Archivos (CIDA);
- Subdirección General de los Archivos Estatales (SGAE).

En el portal se dispone tanto de datos de documentos como del registro de autoridad vinculado a esos documentos. Atendiendo a lo expresado en el propio portal, el alcance de los ficheros de autoridad en el año 2023 supera los 77.000 registros para familias, instituciones, personas, actividades, lugares, conceptos, normas y cargos unipersonales (Portal de Archivos Españoles, n.d.).

Como se puede comprobar y atendiendo a lo expresado por López Cuadrado, la evolución a PARES 2.0 persigue aportar los principios fundamentales para expresar datos con semántica y el sistema de nombrar entidades con URIs fue habilitado como primer paso (López Cuadrado, 2016). Si bien, considerando los elementos de la ya mencionada arquitectura de la Web semántica, los identificadores URI/IRI se ubican en la base de cualquier proyecto de datos enlazados, si bien no es un elemento *per se* único para que los datos archivísticos queden expresados con semántica. En la fuente mencionada los datos están expresados con el lenguaje XML, capa de los datos en las capas de la arquitectura semántica, con el estándar *Encoded Archival Context - Corporate bodies, Persons and Families* (EAC-CPF, RiC-E4 agrupados como *Agent*) (Society of American Archivists, 2011). Por lo tanto, XML es una tecnología que cumple su papel tan orientado a los datos, a falta de considerar la necesidad expresarse con vocabularios que cuente con una semántica que perdure a largo plazo (Dombrowski & Dombrowski, 2010).

Todos los tipos de autoridad quedan integrados en la tabla 1, así como la equivalencia en RiC. Hay un tipo que no está en RiC, pero sí en esta investigación, que es el que se denomina indefinidos. Es un tipo que aglutina a aquellos que no se pueden asignar a ninguna de las autoridades listadas.

Tabla 1. Tipo de autoridad y su RiC-CM.

Tipo de autoridad	RiC-CM (2016)
Cargos unipersonales	RiC-E5: Occupation
Conceptos	RiC-E14: Concept/Thing
Familias	RiC-E4: Agent
Funciones	RiC-E7: Function
Instituciones	RiC-E4: Agent
Lugares	RiC-E13: Place
Normas / Leyes	RiC-E10: Mandate
Personas	RiC-E4: Agent

Este trabajo cumple con un doble objetivo. El primero de ellos es describir los tipos de autoridades que participan como datos en el Portal de Archivos Españoles y hasta finales del año 2023. Tras alcanzarlo, esta investigación identificará la red de relaciones que se establecen entre las autoridades descritas. Con ambos objetivos se delinea el grafo del conocimiento de PARES.

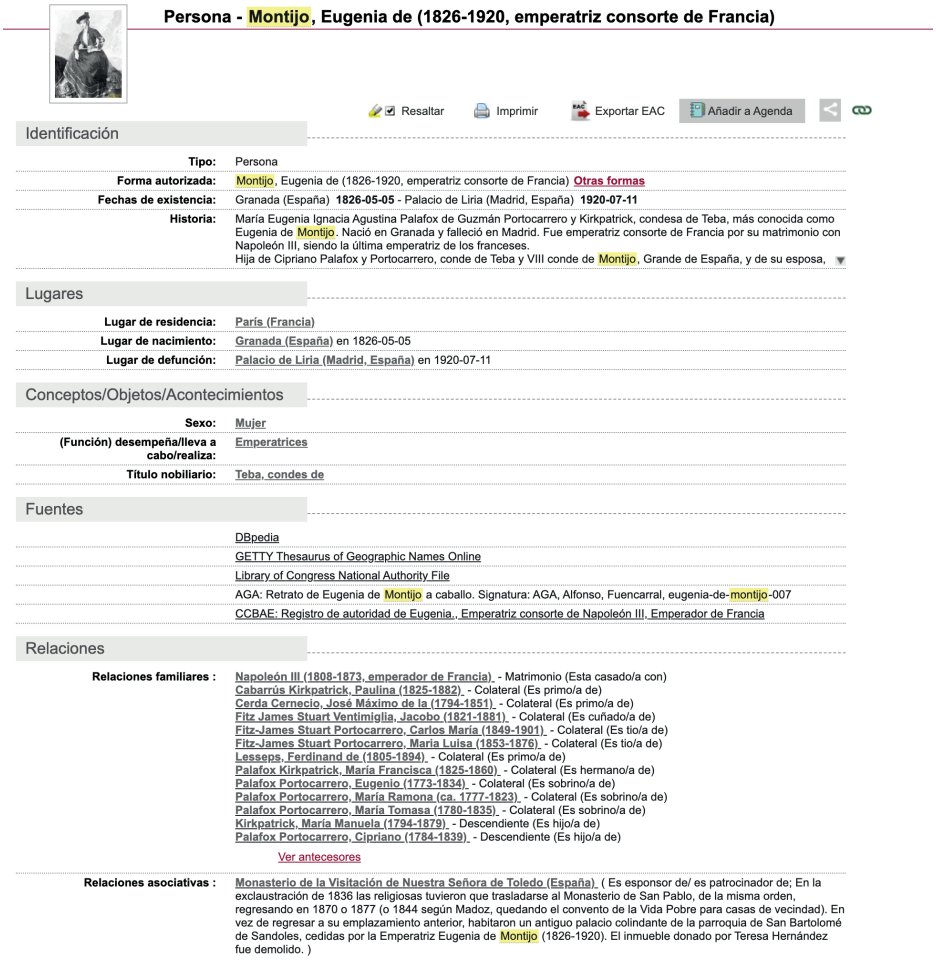
Metodología

La finalidad del programa de *webscraping* fue el buscador de autoridades de PARES. Esta web presenta diversos tipos de autoridad, tales como personas, familias, instituciones, actividades, funciones, lugares, objetos, conceptos, acontecimientos, normas, leyes y cargos unipersonales. Cada una de las entradas de autoridad consta de un identificador numérico que se emplea en la URI que da acceso a su ficha descriptiva. Esto permite una fácil automatización y proceso de extracción de los datos. En concreto el modelo de URI es el que se muestra en la tabla 2.

Tabla 2. URI e identificador de la autoridad Eugenia de Montijo. Fuente: elaboración propia

Persona - Montijo, Eugenia de (1826-1920, emperatriz consorte de Francia)
https://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/**47767**
https://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/exportEAC/**47767**
Identificador: **47767**

Esto permite operar sobre la página HTML o bien con el archivo XML en formato EAC, que también está vinculado. En nuestro caso, el programa se diseñó para operar con el código HTML, debido a que presenta más datos, tales como los documentos relacionados con la autoridad descrita en la ficha.



Persona - Montijo, Eugenia de (1826-1920, emperatriz consorte de Francia)

Resaltar | Imprimir | Exportar EAC | Añadir a Agenda

Identificación

Tipo: Persona

Forma autorizada: Montijo, Eugenia de (1826-1920, emperatriz consorte de Francia) Otras formas

Fechas de existencia: Granada (España) **1826-05-05** - Palacio de Liria (Madrid, España) **1920-07-11**

Historia: María Eugenia Ignacia Agustina Palafox de Guzmán Portocarrero y Kirkpatrick, condesa de Teba, más conocida como Eugenia de Montijo. Nació en Granada y falleció en Madrid. Fue emperatriz consorte de Francia por su matrimonio con Napoleón III, siendo la última emperatriz de los franceses.
Hija de Cipriano Palafox y Portocarrero, conde de Teba y VIII conde de Montijo, Grande de España, y de su esposa,

Lugares

Lugar de residencia: París (Francia)

Lugar de nacimiento: Granada (España) en 1826-05-05

Lugar de defunción: Palacio de Liria (Madrid, España) en 1920-07-11

Conceptos/Objetos/Acontecimientos

Sexo: Mujer

(Función) desempeña/lleva a cabo/realiza: Emperatrices

Título nobiliario: Teba, condes de

Fuentes

DBpedia

GETTY Thesaurus of Geographic Names Online

Library of Congress National Authority File

AGA: Retrato de Eugenia de Montijo a caballo. Signatura: AGA, Alfonso, Fuencarral, eugenia-de-montijo-007

CCBAE: Registro de autoridad de Eugenia,. Emperatriz consorte de Napoleón III. Emperador de Francia

Relaciones

Relaciones familiares :
Napoleón III (1808-1873, emperador de Francia) - Matrimonio (Esta casado/a con)
Cabarrús Kirkpatrick, Paulina (1825-1882) - Colateral (Es primo/a de)
Cerda Cernecio, José Máximo de la (1794-1851) - Colateral (Es primo/a de)
Fitz James Stuart Ventimiglia, Jacobo (1821-1881) - Colateral (Es cuñado/a de)
Fitz-James Stuart Portocarrero, Carlos María (1849-1901) - Colateral (Es tio/a de)
Fitz-James Stuart Portocarrero, María Luisa (1853-1876) - Colateral (Es tio/a de)
Lesseps, Ferdinand de (1805-1894) - Colateral (Es primo/a de)
Palafox Kirkpatrick, María Francisca (1825-1860) - Colateral (Es hermano/a de)
Palafox Portocarrero, Eugenio (1773-1834) - Colateral (Es sobrino/a de)
Palafox Portocarrero, María Ramona (ca. 1777-1823) - Colateral (Es sobrino/a de)
Palafox Portocarrero, María Tomasa (1780-1835) - Colateral (Es sobrino/a de)
Kirkpatrick, María Manuela (1794-1879) - Descendiente (Es hijo/a de)
Palafox Portocarrero, Cipriano (1784-1839) - Descendiente (Es hijo/a de)
Ver antecesores

Relaciones asociativas : Monasterio de la Visitación de Nuestra Señora de Toledo (España) (Es esponsor de/ es patrocinador de; En la exclaustración de 1836 las religiosas tuvieron que trasladarse al Monasterio de San Pablo, de la misma orden, regresando en 1870 o 1877 (o 1844 según Madoz, quedando el convento de la Vida Pobre para casas de vecindad). En vez de regresar a su emplazamiento anterior, habitaron un antiguo palacio colindante de la parroquia de San Bartolomé de Sandoles, cedidas por la Emperatriz Eugenia de Montijo (1826-1920). El inmueble donado por Teresa Hernández fue demolido.)

Figura 1. Ficha de Eugenia de Montijo en PARES.
https://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/47767?nm

El programa de *webscraping* programado en lenguaje PHP, obtiene todos los datos de cada autoridad, en concreto, su tipología, enlace, formas autorizadas, términos preferentes, términos no preferentes, fechas de existencia, lugar de nacimiento, lugar de defunción, lugar de residencia, lugares genéricos, lugares relacionados, latitud, longitud, historia, conceptos y objetos, atribuciones legales, ocupaciones, funciones relacionadas, términos específicos, fuentes de información, relaciones familiares, relaciones asociativas, enlaces externos y documentos relacionados. Adicionalmente la suma de todos los textos de la ficha de autoridad se almacena en campos de indexación, para su recuperación literal y en lenguaje natural, a fin de facilitar su recuperación y consulta en un buscador interno. Esto se consigue con la consulta SQL que se muestra en la tabla 3.

Tabla 3. Consulta SQL y modelo de datos empleado para registrar las autoridades de PARES.

a) [Consulta SQL empleada en el programa de *Webscraping* de autoridades de PARES]

$sql = "**INSERT INTO** autoridades **SET** tipo='**$formType**', enlace='**$urlAA**', formaAutorizada='**$formName**', terminoPreferente='**$prefTerm**', terminoNoPreferente='**$prefNotTerm**', fechasExistencia='**$fechasExistencia**', lugarNacimiento='**$dataLugNac**', lugarDefuncion='**$dataLugDef**', lugarResidencia='**$dataLugRes**', lugarGeneral='**$dataLugGen**', lugaresRelacionados='**$dataLugRel**', latitud='**$latitud**', longitud='**$longitud**', historia='**$dataHist**', conceptosObjetos='**$dataConcept**', atribucionesLegales='**$dataLegal**', ocupaciones='**$dataFunc**', funcionesRelacionadas='**$dataFuncRel**', terminosEspecificos='**$dataTE**', fuentes='**$dataFuentes**', relacionesFamiliares='**$dataFamiRel**', relacionesAsociativas='**$dataAsocRel**', enlacesExternos='**$dataExtLink**', documentosRelacionados='**$dataDocRel**', indexer='**$indexer**', indexerLiteral='**$indexerLiteral**';";

b) [Descripción de datos]

$sql = "**INSERT INTO** autoridades **SET** tipo='**Tipo de autoridad (Persona, Institución, Familia…)**', enlace='**Enlace URI de la autoridad**', formaAutorizada='**Forma autorizada del nombre de la autoridad**', terminoPreferente='**Otras formas aceptadas del nombre**', terminoNoPreferente='**Formas no aceptadas del nombre de la autoridad**', fechasExistencia='**Fechas extremas de la autoridad**', lugarNacimiento='**Lugar y fecha de nacimiento**', lugarDefuncion='**Lugar y fecha de defunción de la autoridad**', lugarResidencia='**Lugar y fechas extremas de los lugares de residencia en los que residió la autoridad**', lugarGeneral='**Lugares genéricos**', lugaresRelacionados='**Lugares relacionados con la autoridad**', latitud='**Latitud**', longitud='**Longitud**', historia='**Historia o nota biográfica completa de la autoridad**', conceptosObjetos='**Conceptos u objetos vinculados a la autoridad**', atribucionesLegales='**Atribuciones legales de la autoridad**', ocupaciones='**Ocupaciones de la aut'ridad**', funcionesRelacionadas='**Funciones de la autoridad**', terminosEspecificos='**Términos específicos vinculados a la autoridad, procedentes del tesauro de PARES**', fuentes='**Reseña de fuentes de información empleadas en la descripción de la autoridad**', relacionesFamiliares='**Relaciones con autoridades familiares**', relacionesAsociativas='**Relaciones con otras autoridades**', enlacesExternos='**Enlaces a recursos y fuentes de información externas**', documentosRelacionados='**Relación de documentos relacionados con la autoridad**', indexer='**Concatenación de todos los campos de texto, limpios, sin palabras vacías, reducción morfológica de términos, normalización del texto, para la facilitar la recuperación de información**', indexerLiteral='**Concatenación de todos los campos de texto, tal como figuran en la ficha de la autoridad, para permitir la búsqueda sensible a mayúsculas, minúsculas, acentos, esto es, literal, de la información de las autoridades.** ';";

c) [Consulta SQL de la autoridad de Eugenia de Montijo. Se muestran datos parciales]

INSERT INTO `autoridades` (`id`, `datepub`, `dateupd`, `tipo`, `enlace`, `formaAutorizada`, `terminoPreferente`, `terminoNoPreferente`, `fechasExistencia`, `lugarNacimiento`, `lugarDefuncion`, `lugarResidencia`, `lugarGeneral`, `lugaresRelacionados`, `latitud`, `longitud`, `historia`, `conceptosObjetos`, `atribucionesLegales`, `ocupaciones`, `funcionesRelacionadas`, `terminosEspecificos`, `fuentes`, `relacionesFamiliares`, `relacionesAsociativas`, `enlacesExternos`, `documentosRelacionados`, `indexer`, `indexerLiteral`) **VALUES** (12572, '2023-11-02 08:51:44', '2024-01-10 10:51:44', 'Persona', 'https://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/47767', 'Montijo, Eugenia de (1826-1920, emperatriz consorte de Francia)', '', '', 'Granada (España)|1826-05-05|Palacio de Liria (Madrid, España)|1920-07-11', 'Granada (España) en 1826-05-05|http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/82294 #@# ', 'Palacio de Liria (Madrid, España) en 1920-07-11 | http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/119859#@#', 'París (Francia) | http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/82613#@#', 'París (Francia) | http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/82613#@#', '', '', '', 'María Eugenia Ignacia Agustina Palafox de Guzmán Portocarrero y Kirkpatrick, condesa de Teba, más conocida como Eugenia de Montijo. Nació en Granada y falleció en Madrid. Fue emperatriz consorte de Francia por su matrimonio con Napoleón III, siendo la última emperatriz de los franceses.**#@#**', 'Mujer|http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/123597#@#Emperatrices|http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/105090#@#Teba, condes de|http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/55727#@#', '', '', '', 'Dbpedia | #@# GETTY Thesaurus of Geographic Names Online| #@# Library of Congress National Authority File | #@# AGA: Retrato de Eugenia de Montijo a caballo. Signatura: AGA, Alfonso, Fuencarral, ortoca-de-montijo-007 | #@# CCBAE: Registro de autoridad de Eugenia., Emperatriz consorte de Napoleón III, Emperador de Francia | #@# ', 'Napoleón III (1808-1873, emperador de Francia) – Matrimonio (Está casado/a con) | http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/47221 | Matrimonio (Esta casado/a con) #@# Cabarrús Kirkpatrick, Paulina (1825-1882) – Colateral (Es primo/a de) | http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/50300 | Colateral (Es primo/a de) #@# Cerda Cernecio, José Máximo de la (1794-1851) – Colateral (Es primo/a de) | http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/157127 | Colateral (Es primo/a de) #@# Fitz James Stuart Ventimiglia, Jacobo (1821-1881) – Colateral (Es cuñado/a de) | http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/109499 | Colateral (Es cuñado/a de) #@# Fitz-James Stuart Portocarrero, Carlos María (1849-1901) – Colateral (Es tío/a de) | http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/49375 | Colateral (Es tío/a de) #@# Fitz-James Stuart Portocarrero, María Luisa (1853-1876) – Colateral (Es tío/a de) | http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/157293 | Colateral (Es tío/a de) #@# ', 'Monasterio de la Visitación de Nuestra Señora de Toledo (España) Es ortocar de/ es patrocinador de; En la exclaustración de 1836 las religiosas tuvieron que trasladarse al Monasterio de San Pablo, de la misma orden, regresando en 1870 o 1877 (o 1844 según Madoz, quedando el convento de la Vida Pobre para casas de vecindad)… | http://pares.mcu.es/ParesBusquedas20/catalogo/autoridad/19777 #@# ', 'Diccionario biográfico RAH|http://dbe.rah.es/ortocarre/13560/ortoca-maria-guzman-y-portocarrero #@# VIAF|http://viaf.org/viaf/39510739 #@# Biblioteca Nacional de España | http://datos.bne.es/resource/XX1090908 #@# Catálogo de autoridades de la Library of Congress | http://id.loc.gov/authorities/n79071077 #@# ', 'No hay Unidades de Descripción asociadas. | #@# Archivo Histórico de la Nobleza | /ParesBusquedas20/catalogo/find?idAut=47767&archivo=3&tipoAsocAut=1&nomAut=Montijo%2C+Eugenia+de+%281826-1920%2C+emperatriz+consorte+de+Francia%29 #@# Archivo General de la Administración | /ParesBusquedas20/catalogo/find?idAut=47767&archivo=4&tipoAsocAut=1&nomAut=Montijo%2C+Eugenia+de+%281826-1920%2C+emperatriz+consorte+de+Francia%29 #@# ', 'persona ortoca emperatriz consorte formas granada palacio liria granada palacio liria ortoca ortoca agustina ortoca ortocarrero condesa conocida…', 'Persona Montijo, Eugenia de (1826-1920, emperatriz consorte de Francia) Granada (España) 1826-05-05 – Palacio de Liria (Madrid, España) 1920-07-11 Granada (España) en 1826-05-05 Palacio de Liria (Madrid, España) en 1920-07-11 París (Francia) París (Francia)…');

En cuanto al método de extracción de datos, se ha creado un código PHP/cURL que permite emplear las funciones de DOM y XPath, de acuerdo con la literatura científica (Bae et al., 2018; Agrawal & Johari, 2019; Chang, 2022). DOM (*Document Object Model*) es una interfaz de programación para documentos HTML y XML. Representa la estructura lógica de documentos y permite que los programas y *scripts* accedan y manipulen el contenido, estructura y estilo de un documento. En términos simples, DOM convierte un documento en un árbol de nodos, donde cada nodo representa una parte del documento (como elementos HTML, atributos, texto, etc.), y los desarrolladores pueden usar métodos y propiedades para acceder y manipular estos nodos (Zhang et al., 2020; Radilova et al., 2022). De esta forma, las páginas web HTML de las autoridades de PARES son adaptadas para la extracción de sus datos. Por otro lado, XPath es un lenguaje de consulta para seleccionar nodos de un documento XML o HTML, basado en su ubicación en el árbol DOM. XPath proporciona una sintaxis compacta y expresiva para navegar y seleccionar partes específicas de un documento XML o HTML

(Gunawan et al., 2019). Esto facilita la navegación, selección y distinción de los distintos campos de información de la autoridad, para su registro en base de datos.

Los pasos que realiza el programa de *webscraping* son los siguientes:

a) definición del bucle general de control, que determina el número total de iteraciones o registros de que dispone PARES;

b) preparación del enlace de consulta, para acceder al recurso de autoridades;

c) comprobación de duplicaciones antes de proceder a la consulta del enlace;

d) recepción del código HTML con la información de la página de autoridades;

e) consulta, selección XPath y preparación de los datos;

f) inserción de los datos en la tabla "autoridades", de acuerdo a lo expresado en el código fuente, disponible en:
https://github.com/manublaz/pares/blob/main/func.pares.php

Análisis

El proceso de extracción se ejecutó el día 31 de octubre de 2023, obteniendo 75.443 registros que se clasifican en torno a las siguientes autoridades (tabla 4).

Tabla 4. Distribución de autoridades de PARES según su tipología. Fuente: elaboración propia

Tipo de autoridad	Número de registros	Tamaño en BD
Cargos unipersonales	358	1.1 MB
Conceptos	10.041	32.7 MB
Familias	702	1.6 MB
Funciones	54	0.7231 MB
Instituciones	9.397	77.4 MB
Lugares	27.004	82.2 MB
Normas / Leyes	439	1.5 MB
Personas	27.447	180.1 MB
Indefinida	1	0.0021 MB

El análisis cuantitativo revela una proporción equitativa entre las autoridades personales y las personas, con más de 27,000 entradas en cada grupo, lo que constituye aproximadamente el 72% del total de entradas de autoridades en PARES. Los Conceptos e Instituciones representan el 13% y el 12%, respectivamente, de los registros recopilados. La participación de otros tipos de autoridades es residual.

Analizando el nivel de interrelación de las autoridades en PARES, observamos que el centro en torno al que pivotan todos los tipos de autoridad son las de tipo personal, quedando latente en el número de relaciones familiares, asociativas, ocupaciones, de conceptos y objetos, lugares relacionados y fuentes. Los lugares son el otro eje fundamental del mapa relacional de autoridades de PARES, ya que cuentan con presencia entre los cargos unipersonales, conceptos, familias, instituciones, normas y leyes y por supuesto las propias autoridades personales, con más de 48.000 relaciones encontradas (tabla 5).

Tabla 5. Número total de relaciones semánticas, disponibles en los datos de autoridades recopilados, incluyendo enlaces a contenidos externos.

Relaciones	Cargos unipersonales	Conceptos	Familias	Funciones	Instituciones	Lugares	Normas / Leyes	Personas
Lugares relacionados	38	71	38	0	3.391	22.931	260	48.635
Conceptos / Objetos	9	0	44	2	2.030	552	26	18.199
Atribuciones Legales	0	0	0	0	8	0	0	0
Ocupaciones	0	0	5	0	104	0	0	16.204
Funciones relacionadas	0	0	0	36	0	0	0	0
Términos específicos	0	787	0	0	0	0	0	0
Relaciones familiares	0	0	92	0	45	0	0	9.614
Fuentes	4	256	1	0	1.656	1.239	13	509
Relaciones asociativas	171	0	68	33	4.322	0	202	12.290

La ratio entre relaciones y autoridades proporciona una visión reveladora sobre la densidad de conexiones entre diferentes tipos de autoridades en un conjunto de datos (tabla 6). En cuanto a lugares relacionados se observa una mayor interrelación entre instituciones, normas, leyes y especialmente personas. Esto sugiere que todas las autoridades personales tendrían una relación con al menos uno o dos lugares. Otro dato significativo es la interrelación de lugares con otros lugares, que alcanza la ratio de 0,85. Esto indicaría que el 85% de los lugares están vinculados con otros lugares o localizaciones. En cuanto a la ratio de conceptos y objetos, la mejor proporción se encuentra con las instituciones y personas y en menor medida con el resto de las autoridades. Lo mismo sucede con el caso de las ocupaciones, principalmente vinculadas a las personas con una ratio de 0,59 ocupaciones por persona, que supondría un 40% de autoridades sin relación de ocupación definida. En cuanto a las relaciones familiares se observa una relación de 3,5 personas por familia, lo cual está dentro de lo normal para la naturaleza de este tipo de relaciones. Significaría que cada núcleo familiar definido en PARES alberga de media entre 3 y 4 personas.

Tabla 6. Ratio de autoridades y sus relaciones. Fuente: elaboración propia

Ratio Relaciones / N Autoridades	Cargos unipersonales	Conceptos	Familias	Funciones	Instituciones	Lugares	Normas / Leyes	Personas
Lugares relacionados	0,106	0,007	0,054	-	0,361	0,850	0,592	1,772
Conceptos / Objetos	0,025	-	0,063	0,037	0,216	0,020	0,059	0,663
Atribuciones Legales	-	-	-	-	0,001	-	-	-
Ocupaciones	-	-	0,007	-	0,011	-	-	0,590
Funciones relacionadas	-	-	-	0,667	-	-	-	-
Términos específicos	-	0,078	-	-	-	-	-	-
Relaciones familiares	-	-	0,131	-	0,005	-	-	0,350
Fuentes	0,011	0,025	0,001	-	0,176	0,046	0,030	0,019
Relaciones asociativas	0,478	-	0,097	0,611	0,460	-	0,460	0,448

Al respecto de las fuentes, la ratio es baja ya que no supera en ninguna autoridad el valor de 0,18 aunque sí está presente en todas las categorías. Esto indica que unas pocas fuentes sirven para describir y apoyar documentalmente las fichas descriptivas de las autoridades en PARES. También son representativas las relaciones asociativas, que tienen ratios que superan los valores de 0,4 en las categorías de cargos unipersonales, funciones, instituciones normas, leyes y personas. Ello indicaría una tasa relacional media con el conjunto de datos.

Discusión

En el contexto de los LAM, los puntos de acceso y las autoridades históricamente han sido una preocupación de la que se han ocupado los profesionales. Gracy (2015) defiende que, en las descripciones archivísticas, el análisis de las frecuencias de los puntos de acceso controlados e incontrolados se sirve de las tecnologías semánticas para un correcto desarrollo de un método analítico enriquecido para personas, familias, organizaciones, nombres geográficos u otras entidades género/forma.

La multitud de relaciones que se generan entre las autoridades revela importantes matices de la información archivística. Así, Niu (2016) plantea que, a la vista de proyectos que han implementado *linked data* para materiales archivísticos, queda confirmado que se mejoran las descripciones y también la recuperación de la información. Por ese motivo, es un gran potencial para un enriquecimiento efectivo e incremento de interoperabilidad de los datos archivísticos.

Durante décadas, la comunidad de archiveros ha valorado, conservado y facilitado el acceso a los documentos de archivo utilizando teorías y métodos archivísticos creados para los documentos de archivo en soporte impreso. Sin embargo, la producción y el consumo de ese corpus documental se ha visto influido por las tendencias sociales e industriales y en métodos centrados en los datos que guardan poca o ninguna relación con los métodos archivísticos más tradicionales (Marciano et al., 2018).

Conclusiones

La principal conclusión es que las relaciones conceptuales dominantes se producen entre los tipos de autoridades, lugares y personas, conceptos e instituciones. Ello demuestra que en el grafo semántico de PARES se da mayor importancia a la geolocalización de ideas y a la comprensión del contexto geográfico de las autoridades. Las relaciones entre personas y familias también son predominantes con un total de 9.614, lo que permite analizar los lazos familiares, la vida social y la genealogía. Se destacan las relaciones asociativas entre personas e instituciones, lo que permitirá realizar minería de datos y descubrir nuevos patrones y conexiones inesperadas entre las autoridades. Ello es compatible con un menor nivel relacional en las atribuciones legales, las funciones relacionadas y términos específicos. Con la descripción de las autoridades y la red de relaciones se ha delineado el grafo del conocimiento de PARES.

Las autoridades menos relacionadas son las funciones y los cargos unipersonales, cuya conexión está intrínsecamente ligada a las instituciones y personas. También se observa que los términos específicos tienen una escasa interrelación con las autoridades, sólo vinculados a los "conceptos". De acuerdo con la estructura de las fichas descriptivas de PARES, los conceptos y los términos específicos sirven para configurar un lenguaje controlado a modo de tesauro, con una organización jerárquica. Sin embargo, no parece haber sido empleado de forma sistemática en la descripción de autoridades familiares, lugares, funciones, personas, normativas o leyes, suponiendo una carencia importante desde el punto de vista temático.

En futuros trabajos de representación semántica de las autoridades de PARES, en el marco de RDF, se deberá comprender el enriquecimiento de los datos recopilados, con otros procedentes de Wikidata y DBpedia. También se plantea el reto de desarrollar un método para la clasificación automática de las autoridades y la asignación de descriptores, que facilite la categorización de los contenidos. De esta forma, se podrá incorporar un *endpoint* que permita la recuperación y generación de grafos del modelo relacional, su representación gráfica e interrogación. También existe la posibilidad de extraer nuevas relaciones de tipo cronológico o temporal, relaciones que precisen el tipo de relación asociativa, relaciones con *N-gramas* presentes en las descripciones y notas biográficas de las autoridades.

Referencias

Agrawal, N., & Johari, S. (2019). A survey on content-based crawling for deep and surface web. *Fifth International Conference on Image Information Processing (ICIIP)* (pp. 491-496). IEEE. https://doi.org/10.1109/ICIIP47207.2019.8985906

APEF (n.d.) *Who we are*. Archives Portal Europe. https://www.archivesportaleurope.net/about-us/who-we-are/

Bae, S. W., Lee, H. D. & Cho, D. (2018). Design and implementation of a web crawler system for collection of structured and unstructured data. *Journal of Korea Multimedia Society*, *21*(2), 199-209. https://doi.org/10.9717/kmms.2018.21.2.199

Chang, Z. (2022). A survey of modern crawler methods. *Proceedings of the 6th International Conference on Control Engineering and Artificial Intelligence* (pp. 21-28). https://doi.org/10.1145/3522749.3523076

CRUE (2017). *Guía Linked Open Data para archivos universitarios*. Grupo de Trabajo Linked Open Data y Archivos Universitarios, CRUE. http://cau.crue.org/wp-content/uploads/GT_9_Gu%C3%ADa_Linked_Open_Data_para_Archivos_Universitarios_2017.pdf

Dombrowski, A., & Dombrowski, Q. (2010). A formal approach to XML semantics: Implications for archive standards. *Proceedings of the International Symposium on XML for the Long Haul: Issues in the Long-Term Preservation of XML*. https://doi.org/10.4242/BalisageVol6.Dombrowski01

Gracy, K. F. (2015). Archival description and linked data: a preliminary study of opportunities and implementation challenges. *Archival Science*, 15, 239-294. https://doi.org/10.1007/s10502-014-9216-2

Guernaccini, F., Mazzini, S., & Bruno, G. (2019). LOD publication in the archival domain: methods and practices. *ODOCH@ CaiSE*, (pp. 15-26). https://ceur-ws.org/Vol-2375/paper2.pdf

Gunawan, R., Rahmatulloh, A., Darmawan, I., & Firdaus, F. (2019). Comparison of web scraping techniques: regular expression, HTML DOM and Xpath. *2018 International Conference on Industrial Enterprise and System Engineering (ICoIESE 2018)*. Atlantis Press (pp. 283-287). https://doi.org/10.2991/icoiese-18.2019.50

Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A. C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J. F., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, *54*(4). https://doi.org/10.1145/3447772

Jacobs, C. T., Avdis, A., Mouradian, S. L., & Piggott, M. D. (2015). Integrating research data management into geographical information systems. *Roceedings of the 5th International Workshop on Semantic Digital Archives (SDA 2015)* (pp. 7–17). http://ceur-ws.org/Vol-1529/paper2.pdf

Koch, I., Freitas, N., Ribeiro, C., Lopes, C. T., & Da Silva, J. R. (2019). Knowledge graph implementation of archival descriptions through CIDOC-CRM. *International conference on theory and practice of digital libraries* (pp. 99-106). Cham: Springer International Publishing.

Llanes-Padrón, D., & Pastor-Sánchez, J.A. (2017). Records in contexts: the road of archives to semantic interoperability. *Program*, 2017, 51(4), 387-405. https://doi.org/10.1108/PROG-03-2017-0021

López Cuadrado, A. M., & Requejo Zalama, J. (2021). Estrategias y modelos de gestión de datos archivísticos. Tábula, 24, 97–111. https://publicaciones.acal.es/tabula/article/view/874

López Cuadrado, A. M. (2016). PARES 2.0: tecnología para mejorar el acceso de los ciudadanos a los documentos y a la información en los Archivos Estatales. En González Cachafeiro, J. (coord.). *Actas de las jornadas 9ª Jornadas archivando: usuarios, retos y oportunidades*. León, 10 y 11 de noviembre (pp. 36-59). ISBN 978-84-617-7452-4

Marciano, R., Lemieux, V., Hedges, M., Esteva, M., Underwood, W., Kurtz, M., & Conrad, M. (2018). Archival records and training in the age of Big Data. In: J. Percell, L. C. Sarin, P. T. Jaeger, & J. C. Bertot (Eds.) *Re-envisioning the MLS: Perspectives on the Future of Library and Information Science Education* (Advances in Librarianship, vol. 44B, pp. 179-199). Emerald Publishing Limited, Leeds. https://doi.org/10.1108/S0065-28302018000044B010

Maynard, D., & Greenwood, M. A. (2012). Large scale semantic annotation, indexing, and search at the national archives. *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC 2012 (pp. 3487–3494). http://www.lrec-conf.org/proceedings/lrec2012/pdf/122_Paper.pdf

Miller, E. (2001). *Semantic Web Layer Cake*. https://www.w3.org/2001/09/06-ecdl/slide17-0.html

Niu, J. (2016). Linked data for archives. *Archivaria*, *82*(1), 83-110. https://archivaria.ca/index.php/archivaria/article/view/13582

O'Reilly, T. (30 de septiembre de 2005). What is Web 2.0: Design patterns and business models for the next generation of software. *O'Reilly*. https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

Portal de Archivos Españoles (n.d.). *Estadísticas de PARES*. https://pares.culturaydeporte.gob.es/estadisticas.html

Radilova, M., Kamencay, P., Hudec, R., Benco, M., & Radil, R. (2022). Tool for parsing important data from web pages. *applied sciences*, *12*(23), 12031. https://doi.org/10.3390/app122312031

Society of American Archivists (2011). *Encoded Archival Context - Corporate bodies, Persons, and Families (EAC-CPF)*. https://www2.archivists.org/node/23669

Vafaie, M., Bruns, O., Pilz, N., Dessí, D. & Sack, H. (2021). Modelling archival hierarchies in practice: Key aspects and lessons learned. *CEUR Workshop Proceedings*, 2981. https://doi.org/10.34657/8006

Zhang, S., Wu, J., & Yang, K. (2020). A webpage segmentation method based on node information entropy of DOM tree. *Journal of Physics: Conference Series*, *1624*(3), 032023. https://doi.org/10.1088/1742-6596/1624/3/032023