

EVENTOS

I JORNADAS DE GESTIÓN DEL
PATRIMONIO BIBLIOGRÁFICO

Toledo, 1-2 de junio de 2017

2017



LA AGREGACIÓN DE METADATOS COMO MODELO DE CONTROL,
VISIBILIDAD Y DISEMINACIÓN DE INFORMACIÓN DIGITAL DE
BIBLIOTECAS, ARCHIVOS Y MUSEOS PATRIMONIALES: EL CASO DE

DIGIHUB



crue

Universidades
Españolas

Red de Bibliotecas
REBIUN

LA AGREGACIÓN DE METADATOS COMO MODELO DE CONTROL, VISIBILIDAD Y DISEMINACIÓN DE INFORMACIÓN DIGITAL DE BIBLIOTECAS, ARCHIVOS Y MUSEOS PATRIMONIALES: EL CASO DE DIGIHUB

METADATA AGGREGATION AS A CONTROL MODEL, VISIBILITY AND DISSEMINATION OF DIGITAL INFORMATION FROM HERITAGE LIBRARIES, ARCHIVES AND MUSEUMS: THE DIGIHUB CASE

Autores:

Xavier Agenjo Bullón. Fundación Ignacio Larramendi. orcid.org/0000-0001-8338-8087 xavier.agenjo@larramendi.es

Maribel Campillejo Suárez. DIGIBÍS. orcid.org/0000-0002-9283-0188
maribel.campillejo@digibis.com

Francisca Hernández Carrascal. DIGIBÍS. orcid.org/0000-0002-2389-0945
francisca.hernandez@digibis.com

Tachi Hernando de Larramendi. DIGIBÍS. orcid.org/0000-0002-9962-8126
tachi@digibis.com

Felisa Matarranz de Antonio. DIGIBÍS. orcid.org/0000-0001-6379-0431
feli.matarranz@digibis.com

Resumen:

Se expone el caso de éxito que supone el protocolo OAI-PMH. Se ilustra con la iniciativa Hispana, agregador nacional español de metadatos de Europeana. Se discute la visión del catálogo colectivo del Grupo de Trabajo de Patrimonio Bibliográfico de las universidades españolas. Se menciona el proyecto ENUMERATE, estadísticas de la digitalización del patrimonio cultural digital europeo. Se valora el uso de tecnologías Linked Open Data en el entorno del Patrimonio Bibliográfico, Archivístico y Museológico. Se propone la creación de un sistema digital basado en el modelo de recolección de metadatos y se aboga por la implantación de un agregador de metadatos conforme a la tecnología Linked Open Data.

Abstract:

The success case of the OAI-PMH protocol is described. It is illustrated with the initiative of Hispana, the national aggregator of metadata. The vision of the Collective Catalogue of the Working Group on Bibliographic Heritage of the Spanish universities is discussed. ENUMERATE statistics digital cultural heritage project is mentioned. The use of Linked Open Data technologies in the context of Archival and Museum Bibliographic Heritage is valued. The creation of a digital system based on the metadata harvesting model is proposed and the implementation of a metadata aggregator according to Linked Open Data technology is advocated.

Palabras clave

Protocolo OAI-PMH; Linked Open Data; Agregadores de metadatos semánticos

Key words

OAI-PMH protocol; Linked Open Data; Semantic metadata aggregator

Desde que el *protocolo OAI-PMH*¹ (*Open Archives Initiative-Protocol for Metadata Harvesting*), para la transmisión de metadatos en Internet, fue definido por Herbert Van de Sompel en el año 1999 su proliferación ha sido espectacular. Los repositorios OAI que corroboran tal éxito están recogidos, junto a la lista de proveedores de datos de la *Open Archive Initiative*², con más de 2.900 repositorios, en las bases de datos OAIster³, ROAR⁴ (*Registry of Open Access Repositories*) u OpenDOAR⁵ (*The Directory of Open Access Repositories*).

Gracias a esta iniciativa informática, en España se han construido importantes bases de datos conforme a este protocolo. Buena cuenta de ello es el proyecto *Hispana*⁶, agregador nacional de contenidos de *Europeana*⁷, que en la fecha en que se escribe

¹ <https://www.openarchives.org/pmh>

² <https://www.openarchives.org/Register/BrowseSites>

³ <http://www.oclc.org/en/oaister.html?urlm=168646>

⁴ <http://roar.eprints.org>

⁵ <http://www.opendoar.org>

⁶ <http://hispana.mcu.es>

⁷ <http://www.europeana.eu>

esta comunicación cuenta con 7.031.735 de objetos digitales. Si se considera que el proyecto empezó en el año 2006 con 25 repositorios OAI y 119.014 objetos digitales⁸, se comprueba que ha habido un incremento de la colección del 5.808'33% en 11 años.

Se da la circunstancia de que el desarrollo de *Europeana* empezó en esos mismos años. Además de lo que ha supuesto desde el punto de vista del trabajo cooperativo, *Europeana* ha contado desde el primer momento en el formato ESE⁹ (*Europeana Semantic Elements*), que es en cierto modo un Dublin Core cualificado, y sobre todo después en EDM¹⁰ (*Europeana Data Model*) (Hernández Carrascal y Juanes Hernández 2013) con un conjunto de reglas, es decir, una ontología, destinadas a publicar en la Web la información en forma de registros RDF¹¹ (*Resource Description Framework*) dentro de la tecnología *Linked Open Data*¹².

DIGIHUB, recolector de metadatos OAI-PMH

Con ese panorama, DIGIBÍS¹³ creó el programa llamado DIGIHUB¹⁴, recolector de metadatos OAI-PMH, actualmente en su versión 4.0. Este sistema se ajusta al *Protocolo OAI-PMH Versión 2.0* del, que utiliza transacciones http para emitir preguntas y obtener respuestas entre un repositorio y un recolector de metadatos, y el desarrollo de sus componentes se ha desarrollado bajo las indicaciones de las premisas *Guidelines for Repository Implementers*, *Guidelines for Harvester Implementers* y *Guidelines for Aggregators, Caches and Proxies*. Y lo hizo para agregar los datos de los repositorios OAI, implementados en una amplia variedad de instituciones, primero en *Hispana* y desde ahí, después, enviarlos a *Europeana* gracias también a la tecnología OAI-PMH, pues *Europeana* a su vez recolecta el repositorio que *Hispana* crea a su vez de forma transparente. De hecho, a día de hoy *Hispana* cuenta, como se ha dicho, con más de siete millones de registros, de

⁸ travesia.mcu.es/portalnb/jspui/bitstream/10421/4765/1/Hispana_SGCB.pdf

⁹ <http://pro.europeana.eu/page/ese-documentation>

¹⁰ <http://pro.europeana.eu/page/edm-documentation>

¹¹ <https://www.w3.org/RDF>

¹² <https://www.w3.org/DesignIssues/LinkedData.html>

¹³ <http://www.digibis.com>

¹⁴ <http://www.digibis.com/software/digihub>

los cuales 2.628.929 están recolectados por *Europeana*, y es el primer agregador nacional, y quinto en el elenco¹⁵, de los agregadores de contenido de *Europeana*.

Siguiendo la senda emprendida por *Europeana* con el uso de la tecnología *Linked Open Data*, el programa DIGIHUB presenta los datos en formato RDF, genera una URI persistente y única para cada objeto, genera URIs dereferenciables, (se puede acceder mediante protocolo http a esa URI y se obtienen datos sobre el objeto) e incluye sistema content-negotiation, mediante el cual el servidor devolverá para las URIs diferentes formatos (HTML o RDF, por ejemplo) según indique el cliente http. Todo ello da buena cuenta de que la agregación de metadatos desde repositorios OAI es un claro caso de estudio. Es decir, la idea de agregar registros a partir de repositorios OAI se ha probado nuevamente.

Agregadores implementados con DIGIHUB

Fruto de esta premisa, Xavier Agenjo expuso, en el *VII Encuentro de Centros Españoles de REDIAL*¹⁶ (*Red Europea de Información y Documentación sobre América Latina*) el día 21 de noviembre de 2013, la posibilidad de contar con un agregador para todas las funciones de la Red. En efecto, la idea fue acogida por la directora de la *Biblioteca de la AECID*¹⁷ (*Agencia Española de Cooperación Internacional para el Desarrollo*), María Araceli García Martín, y con las consabidas dificultades presupuestarias al final se creó un agregador que con el nombre de *Americanae*¹⁸ recoge 168.353 objetos digitales y 16 repositorios en el momento en que se escribe esta comunicación. No solo repositorios españoles, sino de todo el mundo. Tanto europeos, miembros de REDIAL, como americanos, entre ellos los Estados Unidos, Argentina, etc.

También algunas Comunidades Autónomas, siguiendo con la filosofía y el espíritu de la *Ley 16/1985, de 25 de junio, del Patrimonio Histórico Español*¹⁹ otorga a las Comunidades Autónomas la titularidad del patrimonio histórico, han establecido agregadores con las funcionalidades descritas. Es el caso de *Galiciana*, de la

¹⁵ <http://www.europeana.eu/portal/es/explore/sources.html>

¹⁶ <http://www.red-redial.net/america-noticia-6974.html>

¹⁷ <http://www.aecid.es/ES/biblioteca>

¹⁸ <http://www.americanae.es>

¹⁹ https://www.boe.es/diario_boe/txt.php?id=BOE-A-1985-12534

Comunidad Autónoma de Galicia, y *Ciconia*, de la Comunidad Autónoma de Extremadura. Ninguna de ellas en producción cuando se está escribiendo esta comunicación, pero están ya implementadas y a falta de pequeños detalles se publicarán en la Web, en fecha muy próxima.

En *Galiciana* y *Ciconia* no solo se agregan registros de bibliotecas desde repositorios OAI, sino también de archivos. En efecto, ya se han agregado centenares de miles de registros. En el caso de *Galiciana*, el repositorio *Archivo Dixital de Galicia*²⁰ (que también usa las aplicaciones DIGIBIB²¹ y DIGIARCH²² de DIGIBÍS), por ejemplo, aporta 34 archivos diferentes combinándose registros bibliográficos y archivísticos. Los primeros se generan a partir de registros MARC 21²³ y los segundos de EAD²⁴ (*Encoded Archival Description*), en su versión EAD3 (Agenjo Bullón y Hernández Carrascal 2016). Los registros son generados de forma dinámica y transparente para el bibliotecario o el archivero desde la aplicación, DIGIBIB o DIGIARCH, al repositorio OAI y en ambos casos se cuenta con una extensión EDM. En el caso de la *Biblioteca Dixital de Galicia*²⁵ ha tenido procesos de enriquecimiento semántico y, de hecho, cuenta con un SPARQL endpoint²⁶. Y lo más importante para los fines de esta comunicación es que *Galiciana* recolecta repositorios OAI de instituciones culturales, como la *Real Academia Galega*²⁷, y repositorios OAI de universidades gallegas, como el repositorio *Minerva*²⁸ de la Universidad de Santiago de Compostela o el repositorio *RUC*²⁹ de la de la Coruña.

El Grupo de Patrimonio Bibliográfico de las universidades españolas

Por otro lado, Las universidades españolas cuentan con un grupo de trabajo denominado *Patrimonio Bibliográfico*³⁰ (de la *Red de Bibliotecas Universitarias Españolas*, REBIUN). Ahora bien, al igual que ocurre con las bibliotecas virtuales o digitales de las Comunidades Autónomas o de instituciones autónomas, como reales

²⁰ <http://archivo.galiciana.gal/arpadweb/gl/archivos/centrosArchivo.do>

²¹ <http://www.digibis.com/software/digibib>

²² <http://www.digibis.com/software/digiarch>

²³ <http://www.loc.gov/marc/>

²⁴ <http://www.loc.gov/ead>

²⁵ <http://biblioteca.galiciana.gal>

²⁶ <https://www.w3.org/wiki/SparqlEndpoints>

²⁷ <http://academia.gal/>

²⁸ <https://minerva.usc.es>

²⁹ <http://ruc.udc.es>

³⁰ <http://www.rebiun.org/patrimonioBibliografico/Paginas/default.aspx>

academias, ateneos, etc., se ha aprovechado la circunstancia de que disponen de repositorios OAI para que sus colecciones digitales sean recolectadas por *Hispana*.

En el caso de las bibliotecas de las universidades se cuenta con el agregador *Recolecta*³¹, pero este agregador no cuenta con un mecanismo de agregación ágil, como se puede ver por las estadísticas. Así, *Recolecta* cuenta solamente con 980.000 documentos³². Ahora bien, *Hispana* también recolecta repositorios de universidades. Se trata de 95 repositorios OAI que agregan más de tres millones cuatrocientos mil registros por *Hispana*. Lamentablemente en un porcentaje muy importante no cuentan con EDM, incluso tampoco disponen de ESE, limitándose al preceptivo Dublin Core. Por tanto, los repositorios de las bibliotecas universitarias que son agregados por ese procedimiento a *Europeana* aportan 279.987.

Es importante recalcar que las revistas editadas con el programa OJS³³ (*Open Journal Systems*) cuentan con repositorios OAI y, aunque muchas veces los propios editores no conocen esa posibilidad, lo que es lástima, puesto que da una información riquísima ya que esa funcionalidad permite agregar a los recolectores todos los artículos publicados en esas revistas. Por ejemplo, es el caso de los *Cuadernos de Estudios Gallegos*³⁴, recolectados en Galiciana, aunque aún no está accesible o el de *AIBR, Revista de Antropología Iberoamericana*³⁵, en Americanae. Es decir que, resaltando la ventaja del protocolo OAI-PMH, es muy factible agregar los metadatos generados por todo tipo de instituciones de forma muy fácil.

Estadísticas de digitalización del patrimonio cultural digital en Europa

La Comisión Europea mide el esfuerzo de la digitalización del patrimonio cultural digital a través de *Europeana*. Cada dos años se evalúa el grado de digitalización del patrimonio y se concluye qué ha sido digitalizado cuando la información está disponible en *Europeana*. Desde el año 2015, *Europeana* cuenta, entre su versión 3, con el proyecto ENUMERATE³⁶ que mide el esfuerzo de digitalización en Europa a partir de una serie de encuestas realizadas a todos los estados miembros. Los

³¹ <https://recolecta.fecyt.es>

³² <https://buscador.recolecta.fecyt.es>

³³ <https://pkp.sfu.ca/ojs>

³⁴ <http://estudiosgallegos.revistas.csic.es/index.php/estudiosgallegos>

³⁵ <http://www.aibr.org/antropologia/netesp/>

³⁶ <http://pro.europeana.eu/structure/enumerate>

resultados de esas encuestas se publican en la plataforma digital ENUMERATE Data Platform³⁷ –implementada por DIGIBÍS–. En cualquier caso, se cuenta con una información más detallada en el texto *El observatorio de la digitalización en Europa*³⁸, cuyos dos autores, Xavier Agenjo y Maribel Campillejo, firman también la presente comunicación (Agenjo Bullón y Campillejo Suárez 2015).

De esta forma, la Comisión Europea ha publicado nuevamente (junio 2016) el informe *Digitisation, online accessibility and digital preservation*³⁹, donde ENUMERATE ha sido la fuente de información sobre el estado de la digitalización en Europa. En lo que se refiere a España, los indicadores están basados en la agregación de contenidos en el agregador nacional *Hispana*, donde se da la información pertinente tanto de digitalizaciones como de metadatos, ESE y EDM, puesto que como dice el informe, *Hispana* requiere a los repositorios que cuenten con copyright, etc.

Following the requirements from Europeana, HISPANA is asking data providers to state copyright conditions in their metadata. If their metadata don't have such statement, HISPANA places in the EDM or ESE registry the creative commons public domain statement. Anyway, open access to a diffusion copy (lower resolution) through these digital libraries is commonly widespread. The public use of such images and/or need of higher resolution images are normally charged, but not because of intellectual property issues, but for covering the service of image provision. Besides for cultural institutions this income (especially for museums) might be a substantial way of complementing the yearly shortening budgets.

Así, cuando se consulta *Europeana*, desde el repositorio *Gredos*⁴⁰ de la Universidad de Salamanca (que cuenta con EDM) hasta el repositorio de la Universidad del CEU⁴¹, las universidades logran dar una visibilidad mayor, puesto que desde la agregación de *Hispana* y después la agregación de *Europeana* los datos están

³⁷ <http://www.egmus.eu/?id=391>

³⁸ <http://eprints.ucm.es/34723/>

³⁹ http://ec.europa.eu/information_society/newsroom/image/document/2016-43/2013-2015_progress_report_18528.pdf

⁴⁰ <https://gredos.usal.es/jspui>

⁴¹ <http://dspace.ceu.es/>

disponibles en muchas visualizaciones en diferentes formatos, Dublin Core, marc, ESE, EDM, etc. En el caso de DIGIBÍS, los registros de las bibliotecas digitales o virtuales que cuentan con el programa DIGIBIB y el repositorio OAI consiguiente son también agregados por OAIster⁴² y se logra una mayor visibilidad. Es posible también que *Americanae* agregue en su día las colecciones digitales de esas universidades que sirvan a sus propósitos americanistas, etc.

La agregación de metadatos a través de los recolectores. Un caso de éxito

Está más que probada la agregación de los metadatos como modelo de control, visibilidad y diseminación de bibliotecas, archivos y museos está muy probada. Por lo tanto, sería deseable que los esfuerzos del Grupo de Patrimonio Bibliográfico de REBIUN se complementaran con las técnicas basadas en repositorios OAI y la agregación de datos con programas creados con ese fin. Como es lógico, nosotros recomendamos nuestro software DIGIHUB.

El *Catálogo de REBIUN*⁴³ cuenta, fruto de los esfuerzos de los bibliotecarios, con un subcatálogo de Fondo Antigo de todas las bibliotecas de las universidades. Sin embargo, este catálogo no cuenta con ninguna de las posibles funcionalidades que le daría la publicación en un repositorio OAI. En este sentido, el programa DIGIHUB que es un sistema ideal para redes, consorcios, catálogos colectivos y cualquier otro tipo de asociación de bibliotecas, archivos o museos interesadas en la recolección de los objetos digitales de patrimonio o de investigación de una región o de un país, permitiría a REBIUN recolectar los recursos digitales proporcionados por archivos, bibliotecas y museos, generando un punto único de consulta a toda esa producción digital recolectada y un repositorio OAI-PMH que a su vez facilitaría la agregación de los recursos digitales a otros servicios de recolección.

Por lo tanto, ese catálogo colectivo y, sobre todo, las digitalizaciones asociadas, podría ser recolectado por *Hispana*, diferenciando los registros catalográficos y los objetos digitales, es decir, aquellos registros que cuenten con una digitalización asociada, una imagen, etc., y desde allí se podrían incorporar en su caso a *Europeana*. Una forma mucho más eficiente sería crear un agregador propio,

⁴² <http://oaister.worldcat.org>

⁴³ <http://rebiun.baratz.es/rebiun/>

análogo a *Hispana*, a *Americanae*, a *Galiciana*, etc., puesto que desde allí se podría realizar el enriquecimiento semántico de los registros. Y sería lo mismo para el patrimonio archivístico y museológico de las universidades como prueban los dos repositorios de universidades recolectados en el proyecto *Galiciana* o el repositorio OAI del *Museo Felix Cañada*⁴⁴, adscrito a la Escuela Técnica Superior de Ingenieros de Minas y Energía de la Universidad Politécnica de Madrid, implementado sobre el programa DIGIMUS⁴⁵ también de DIGIBÍS, que cuenta con EDM y enriquecimiento semántico, entre ellos los propios tesauros de las Subdirección General de Museos Estatales⁴⁶ y muy importante el *ATT (Art & Architecture Thesaurus)* el *ULAM (The Union List of Artist Names)*, ambos vocabularios de la de la *Paul Getty Foundation*⁴⁷.

Invitación y reto

Recientemente, la *Fundación Ignacio Larramendi* ha presentado la *Biblioteca Virtual de Viajeros Científicos Ilustrados*⁴⁸, basada en la implementación del software DIGIBIB, en su versión 10, y se ha aprovechado para diseñar una interfaz nueva no solo para esa biblioteca virtual, sino para toda la base. Desde el punto de vista digital, se han efectuado muchas mejoras en esta biblioteca, que están descritas en la *Nota a la edición digital de la biblioteca*⁴⁹. Entre ellas se puede destacar la Línea del tiempo, las Ocupaciones y las Formaciones, que están codificadas en los campos 372 y 374, las Fechas de lugares de nacimiento de los autores, codificadas en el campo 331, y lo más importante se han asignado los vocabularios de valores en los campos 024. Pues bien, todas estas funcionalidades están basadas en un desarrollo de DIGIBÍS para, a partir de registros MARC 21 extraer la información conforme a todas esos vocabularios de valores, como *VIAF*⁵⁰, *Wikidata*⁵¹; vocabularios de valores de bibliotecas y catálogos colectivos del entorno cultural de la *Biblioteca Virtual de Polígrafos*⁵², como por ejemplo la *Biblioteca Nacional de*

⁴⁴ <http://museovirtualefexcañada.digibis.com/es>

⁴⁵ <http://www.digibis.com/software/digimus>

⁴⁶ <http://tesauros.mecd.es/tesauros>

⁴⁷ <http://www.getty.edu/research/tools/vocabularies/>

⁴⁸ <http://www.larramendi.es/vcillustrados/i18n/micrositios/inicio.do>

⁴⁹ http://www.larramendi.es/vcillustrados/es/cms/elemento.do?id=ms%2Fvcillustrados%2Fpaginas%2FNota_a_esta_edicion_digital.html

⁵⁰ <https://viaf.org/>

⁵¹ https://www.wikidata.org/wiki/Wikidata:Main_Page

⁵² <http://www.larramendi.es/es/inicio/inicio.do>

España⁵³ (BNE), la *Deutsche National Bibliothek*⁵⁴ (DNB), *Bibliothèque nationale de France*⁵⁵ (BNF), la *Library of Congress*⁵⁶ (LC), el *Catalogue du Système Universitaire de Documentation*⁵⁷ (SUDOC) o el *WorldCat*⁵⁸; y a las *DBpedias*⁵⁹ en los idiomas castellano, catalán, gallego, portugués y vasco.

¿Por qué el Grupo de Patrimonio Bibliográfico de las universidades españolas no se plantea comenzar un camino similar con tecnología OAI-PMH, la publicación de los registros y de los objetos digitales en RDF conforme a la ontología EDM, la asignación de los vocabularios de valores, la agregación a partir de *Hispana*, *Europeana* y *OAIster*, en la Web Semántica? DIGIBÍS está para ello a su completa disposición.

Bibliografía

- AGENJO BULLÓN, Xavier; CAMPILLEJO SUÁREZ, Maribel. El observatorio de la digitalización en Europa. En *VII Encuentro Ibérico EDICIC 2015*: 16 y 17 de noviembre. Madrid. [Fecha de consulta: 12/04/2017] ISBN 978-84-608-3330-7. Disponible en <http://eprints.ucm.es/34723/>
- AGENJO BULLÓN, Xavier; HERNÁNDEZ CARRASCAL, Francisca. EAD3: registros de autoridad, datos abiertos y linked open data. En *Tabula: revista de Archivos de Castilla y León*, 2016, n. 19, p. 311-350. [Fecha de consulta: 12/04/2017] ISSN 1132-6506. Disponible en <http://publicaciones.acal.es/index.php/tabula/article/view/632>
- HERNÁNDEZ CARRASCAL, Francisca; JUANES HERNÁNDEZ, César. La participación española en Europeana: los modelos de datos. En *Boletín de ANABAD*, 2013, LXIII, n. 3, p. 539-557 [Fecha de consulta: 12/04/2017]. ISSN 2010-4164. Disponible en <http://www.digibis.com/images/PDF/anabad-2013-3-participacionespanola.pdf>

⁵³ <http://datos.bne.es/inicio.html>

⁵⁴ http://www.dnb.de/DE/Home/home_node.html

⁵⁵ <http://data.bnf.fr/>

⁵⁶ <http://id.loc.gov/index.html>

⁵⁷ <https://www.idref.fr/autorites/autorites.html>

⁵⁸ <http://experimental.worldcat.org/fast/>

⁵⁹ <http://wiki.dbpedia.org/>

- MARTÍNEZ CONDE, María Luisa. De Hispana a Europea: la integración de todo el patrimonio cultural digital en un sistema cooperativo. En *Boletín de Anabad*. 2014, LXIV, n. 4, p. 400-413 [Fecha de consulta: 12/04/2017]. ISSN 0210-4164. Disponible en <http://www.anabad.org/publicaciones/boletin/3745-2014-num-4-octubre-diciembre-boletin-de-anabad>

