SINCRONIZACIÓN EN EL CONTROL DE AUTORIDADES REPOSITORIO/CRIS DE LA UNIVERSIDAD CARLOS III DE MADRID

Ana Poveda Poveda, Victoria Rasero Merino,
Belén Fernández-del-Pino Torres y José Luis Martín Muñoz

RESUMEN: Se presenta el flujo de trabajo y las tareas realizadas para dotar el repositorio institucional de la Universidad Carlos III de Madrid, *e-Archivo*, de un módulo de Autoridades interoperable con el CRIS institucional. El objetivo es avanzar en la calidad de los metadatos mediante la normalización de la forma de nombre de los autores investigadores, lo que conlleva la agrupación de su producción científica en un único punto de acceso, además de incorporar como valor añadido la conexión con el identificador ORCID y otros perfiles de autor externos, que completan la trayectoria investigadora de cada autor.

Los agentes implicados son los administradores de dichas plataformas, procedentes de los Servicios de Investigación y Biblioteca de la Universidad Carlos III de Madrid en colaboración con la empresa *Arvo Consultores y Tecnología*.

Palabras clave: CRIS; repositorio; autoridades; identificador; investigación; código abierto; acceso abierto; calidad.

ABSTRACT: We present the workflow and the tasks carried out to implement the Authorities control into University Carlos III of Madrid institutional repository, *e-Archivo*, interoperable with the institutional CRIS. The aim is to improve metadata quality by normalizing researchers name, which results in a single access point for each one. In addition, and as an added value,

facilitates the connection with external author profiles, as ORCID, which complete the researcher trajectory.

The agents involved are the administrators of both platforms, coming from the Research and Library Services at University Carlos III of Madrid, as well as *Arvo Consultores y Tecnología* company.

Keywords: CRIS; repository; author control; identificator; research; open source; open access; quality.

1. INTRODUCCIÓN

Cuando en 2006 se puso en marcha el repositorio institucional de la Universidad Carlos III de Madrid (UC3M), *e-Archivo*, se pretendían cubrir varios objetivos:

- integrar, conservar y preservar la producción intelectual de la Universidad;
- aumentar la visibilidad de la obra, del autor y de la Universidad;
- aumentar el impacto de la producción científica disponible en red;
- proporcionar acceso a la información sin restricciones.

Con el paso de los años, se han unido otros objetivos como el de asegurar el cumplimiento de la *Ley* 14/2011, *de* 1 *de junio*, *de la Ciencia*, *la Tecnología y la Innovación*, del *Real Decreto* 99/2011, *de* 28 *de enero por el que se regulan las enseñanzas oficiales de doctorado* (art. 14.5) y de la normativa de las agencias financiadoras de la investigación: Séptimo Programa Marco, Horizonte 2020, Secretaría de Estado de Investigación, Comunidad de Madrid, etc.

La vinculación del CRIS y del repositorio se remonta al *Proyecto de integración entre la base de datos de la actividad investigadora o* CRIS *(Universitas* XXI – *Módulo* IAI*) y el repositorio institucional de la* UC3M *(e-Archivo)*[1], llevado a cabo durante el año 2012 y cuya prioridad era proporcionar un único modelo de autoarchivo de resultados y datos de investigación en el repositorio institucional, acorde con la vía verde de acceso abierto.

Actualmente, *e-Archivo* ofrece una colección de más de 20.000 documentos en acceso abierto que abarca tesis doctorales, revistas editadas por la UC3M, actas de congresos, artículos, libros y capítulos, documentos de trabajo, preprints, conjuntos de datos, etc. Esto conlleva la existencia de un índice de más de 20.000 entradas de autor, cuyo control y depuración de

[1] Galardonado con el Premio de Excelencia 2013 – Modalidad Personal de Administración y Servicios, otorgado por el Consejo Social de la Universidad Carlos III de Madrid http://portal.uc3m.es/portal/page/portal/conocenos/premios_excelencia_2013

posibles duplicados y variantes de nombre suponía un paso más para mejorar la calidad del repositorio que, entre otros aspectos, debía contemplar la normalización de las formas de nombre de los investigadores de la UC3M con el objeto de relacionar sin ambigüedad la producción científica con sus respectivos autores.

De esta manera, se ha implantado el módulo de Autoridades de DSpace para efectuar la desambiguación (unificación y normalización de formas de nombre) de autores, otorgar mayor precisión en la recuperación de la información y relacionar a cada autor con un identificador digital único.

Así pues, cuando un usuario llega a un documento específico y quiere consultar más documentos del mismo autor, al hacer clic en el nombre del autor obtiene como resultado todos los ítems asociados y no solo los que están bajo una variante de nombre o ítems de otro autor con la misma forma de nombre, tal y como ocurría previamente.

Por añadidura, se ha procedido a asociar cada registro de autoridad al identificador ORCID, así como a otros identificadores reconocidos en el ámbito científico como ResearchID y ScopusID.

En este texto queremos presentar el protocolo establecido de sincronización de la base de datos de Autoridades en DSpace con la del CRIS.

2. MATERIALES Y METODOLOGÍA

2.1. ANTECEDENTES

Como en toda gran institución, los proyectos exitosos conllevan una estrecha colaboración entre distintos servicios. En nuestro caso, representamos al Servicio de Investigación y al de Biblioteca y, entre los proyectos realizados en colaboración, está el anteriormente citado de integración entre la base de datos de la actividad investigadora o CRIS (Dos en una: se integran las bases de datos de Información de la Actividad Investigadora y el Archivo Abierto. En: Digital 3 (24), p. 22. Recuperado en: http://hdl.handle.net/10016/14717), que supuso el paso necesario para este nuevo proyecto y cuyo flujo de trabajo se muestra en la Figura 1.

Figura 1. Integración entre ambos sistemas

Otro proyecto llevado a cabo en colaboración entre el Servicio de Investigación y el de Biblioteca, que reforzó el recién mencionado e impulsó al que aquí presentamos, fue la implantación de ORCID en la Universidad Carlos III de Madrid, por la que se procedió a la dotación de un identificador digital único para cada investigador en activo. En realidad, en el informe sobre la implantación del ORCID (FERNÁNDEZ-DEL-PINO, RASERO, & TOLEDO, 2014) ya adelantábamos:

> En el momento de redactar este documento, nuevos proyectos relacionados van poniéndose en marcha:
>
> [...]
>
> – Preparar el repositorio institucional (*e-Archivo*), basado en el *software* DSpace, para poder incluir el campo ORCID y que éste sea recuperable.

Efectivamente, se constataba la necesidad de articular un mecanismo de control de autoridades de personas en *e-Archivo*, que por entonces superaba los 20.000 autores, muchos de ellos duplicados, incompletos, con nombres idénticos para distintos autores, y otros problemas que conllevaba la falta de control.

2.2. Fases

2.2.1. *Implantación del módulo de Autoridades en e-Archivo. Contratación de la empresa Arvo Consultores y Tecnología Consultores y Tecnología*

Tras analizar las posibilidades que ofrecía el programa DSpace respecto al control de Autoridades (Luyten & Wood, 2015), se vio la necesidad de contratar a la empresa *Arvo Consultores y Tecnología* para la implantación del módulo de Autoridades desarrollado por ellos sobre la base del propio módulo de DSpace. La explicación de este desarrollo se presenta en su *blog Hablando de DSpace* (Román, 4 marzo 2014). escrito por la citada compañía:

> El Control de Autoridades o Authority Control es una de las piezas clave a disposición de los responsables de Repositorios Digitales para mejorar la calidad de contenidos y posibilitar la interoperabilidad entre repositorios.

- Una autoridad es un conjunto de valores controlados para un dominio determinado, estando cada valor único identificado por una clave (clave de autoridad).
- Un registro de autoridad es la información asociada con cada uno de los valores en una autoridad (incluyendo variaciones de deletreo, valores equivalentes y/o alternativos, etc.).
- Una clave de autoridad es un identificador opaco y persistente correspondiente a un registro de autoridad.

> En la práctica habitual, un registro de autoridad (de nombres de autor, por ejemplo), contiene la forma autorizada del nombre del autor, establecida por la institución normalizadora como forma preferida para visualizar en sus sistemas, así como las formas variantes del nombre y nombres relacionados. Además, el registro de autoridad puede contener información relativa a la persona, representada por el punto de acceso), así como a las relaciones entre esa persona y otras entidades relacionadas, información para identificar las reglas de acuerdo con que se establecieron valores controlados, las fuentes consultadas, la agencia de catalogación encargada de establecer la normalización y la agencia responsable de establecer las formas preferidas del nombre.

Es preciso resaltar que este desarrollo permite la gestión propia de la base de Autoridades del repositorio. Así pues, las tareas de gestión que el administrador del repositorio puede realizar se basan en 3 pasos:

- exportación de la base de datos a formato csv.
- modificación de la información.
- importación del fichero csv actualizado.

Tras realizar varios test en el entorno de pruebas, en abril de 2017, se implantó el control de Autoridades en producción.

El control de Autoridades en *e-Archivo* se realiza exclusivamente sobre los metadatos contributor.author y contributor.advisor, sin perjuicio de que en un futuro se pueda aplicar el control a otros metadatos.

2.2.2. *Volcado de los datos necesarios de los investigadores desde el CRIS para su cotejo y depuración (nombre, apellidos, id interno y ORCID)*

El CRIS de la Universidad, gestionado por el Servicio de Investigación, contiene los datos de todos los investigadores de la universidad procedentes del sistema de gestión de Recursos Humanos. A efectos de volcar dichos datos en el repositorio, se diseñó en el sistema del CRIS una consulta que recuperara los nombres de los investigadores en activo, asociados a su código ORCID y al identificador interno, contenidos en un fichero csv.

Como resultado se obtuvo un listado con el aspecto mostrado en la tabla de la Figura 2, que contiene los siguientes campos:

- ID: Identificador propio del CRIS, que pasa a ser el Identificador clave en la tabla de Autoridades de e-Archivo. Destacamos la importancia de aprovechar este código ya existente que facilita la sincronización entre ambos sistemas.
- Nombre.
- Apellidos.
- Departamento.
- Correo electrónico.
- ORCID.
- ID-DIALNET (en caso que esté registrado en el CRIS).
- ID-SCOPUS (en caso que esté registrado en el CRIS).
- RESEARCHER-ID (en caso que esté registrado en el CRIS).

Figura 2. Identificadores de autor

	A	B	C	D	E	F	G	H	I
1	ID	NOMBRE	APELLIDOS	DEPARTAMENTO	EMAIL	ORCID	DIALNET	SCOPUS	RESEARCHERID
2	99999	José	Pérez García	TEORÍA DE LA SEÑAL Y COMUNICACIONES	jpg@ing.uc3m.es	0000-0000-0000-0000		9999999999	A-0000-2012

2.2.3. *Carga de autores en la base de datos de Autoridades en e-Archivo*

Tras recibir el listado generado por el Servicio de investigación se procedió a realizar las siguientes tareas:

2.2.3.1. Corrección de los nombres y apellidos (originalmente en mayúsculas y sin tildes).

2.2.3.2. Importación del fichero csv de autores, una vez depurado, a la tabla de Autoridades de DSpace, a través de la interfaz de administración. Tal y como se muestra en la Figura 3.

2.2.3.3. Tarea automática de curación para asociar los registros existentes en el repositorio (más de 20.000 ítems) a la autoridad correspondiente.

Figura 3. Importación/exportación de datos

En este caso, la tarea de curación automática exige una coincidencia exacta de la forma de nombre de autor para su validación. Es decir, si el nombre de autor coincide de manera exacta con el existente en la base de datos de autores, este se valida, pasando a utilizar la clave indicada en la misma y no la automática asignada por la herramienta. En caso contrario no se valida.

2.2.3.4. Tarea manual de actualización:

2.2.3.4.1. Por autores: en el repositorio seleccionamos un autor y editamos los metadatos de sus ítems, uno por uno, buscando y escogiendo el registro de autoridad correspondiente.

2.2.3.4.2. Por colecciones: desde la interfaz de administración utilizamos la funcionalidad de exportación/importación de metadatos de una comunidad o colección (opciones «exportar metadatos» e «importar metadatos»), editamos el csv (modificando tanto el nombre con la forma que aparece en la base de Autoridades como la clave de autoridad validada) y volvemos a importar el csv actualizado. Así se aplicarían los cambios masivos conservando la integridad de los índices.

2.2.3.4.3. Programar la tarea de curación automática diariamente, como se refleja en la Figura 4.

Figura 4. Tareas de curación del sistema

2.2.4. *Acciones que se deben planificar para el mantenimiento del sistema de Autoridades de manera sincronizada*:

2.2.4.1. Configuración de un servicio de alerta de nuevas incorporaciones de investigadores en el CRIS. Se crea un proceso automático que se lanza semanalmente con el fin de obtener los datos de los nuevos investigadores que se van dando de alta en el sistema, para agregarlos a la base de Autoridades del repositorio. De esta forma el CRIS y el repositorio están sincronizados, vinculando los datos de los investigadores entre ambos sistemas gracias al identificador interno.

2.2.4.2. Solicitud de ORCID para los nuevos investigadores.

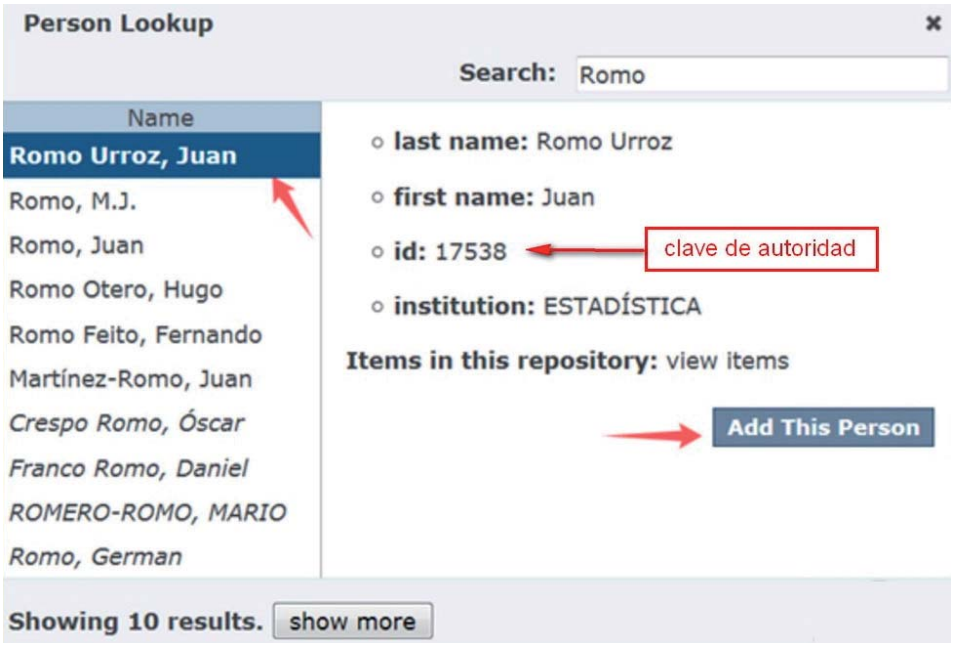
2.2.4.3. Curación en bloques (colecciones o autores) de los nuevos datos recibidos desde el CRIS.

2.2.4.4. Corrección de posibles errores existentes en la base de Autoridades.

2.2.4.5. En algún caso, el investigador desea que figure la forma de su nombre coincidiendo con la firma que utiliza en sus publicaciones, que a su vez aparece en diferentes bases de datos que indexan sus obras. Será necesario, pues, asociar la forma de nombre requerida a la clave de autoridad.

La unificación de nombres de autor en el módulo de Autoridades está garantizada mediante la asociación de una forma de nombre normalizada, que en nuestro caso se trata del nombre oficial registrado en la base de datos de Recursos Humanos, con una clave de autoridad diferenciada, como se aprecia en la Figura 5.

Figura 5. Clave de autoridad



3. RESULTADOS

Con este proyecto se consigue un avance significativo en la calidad de los metadatos en *e-Archivo*, a la vez que se obtiene una gran precisión en la atribución de la autoría. Se incorpora como valor añadido la conexión con perfiles de autor externos, que completan la trayectoria investigadora de cada autor.

Como resultado del trabajo realizado, visualizando la interfaz del administrador de DSpace, se distinguen tres posibilidades:

a) Los autores pertenecientes a la Universidad (con clave de autoridad específica y forma normalizada) aparecen en negrita.
b) Los autores existentes en e-Archivo, pero no pertenecientes a la base de datos de autores de la Institución aparecen en letra estándar (la herramienta les asigna una clave de autoridad automática del tipo 5cf62414-43db-470e-ad48-6acbb30b2bf8).
c) Los autores de la base de datos ORCID aparecen en cursiva (la herramienta les asigna una clave de autoridad automática del tipo 5cf62414-43db-470e-ad48-6acbb30b2bf8).

Por su parte, en la versión pública de *e-Archivo*, estas diferencias tipográficas resuelven los valores controlados por autoridad con un símbolo y su correspondiente enlace a la página existente en ORCID (y a otros identificadores que iremos incorporando: ResearchID, ScopusID, Dialnet, etc.), mostrándose como ilustra la Figura 6.

Figura 6. Símbolos de identificación

Los registros ya creados en *e-Archivo* en el momento en el que se implantó el módulo de Autoridades, tienen que ser revisados para asignar la clave de autoridad validada a los autores UC3M, en los metadatos dc.contributor.author y dc.contributor.advisor. Esta tarea llevará cierto tiempo, pues no parece posible su automatización. Por ello, tras la puesta en marcha, hay que dedicar un periodo de tiempo a la revisión del índice de autores para detectar entradas que no estén validadas.

4. DISCUSIÓN Y CONCLUSIONES

Con este proyecto se pretende, además de lo ya expresado, seguir ofreciendo un alto grado de calidad en la gestión de *e-Archivo*, tal y como se constata en los siguientes logros:

- Visibilidad nacional de la investigación realizada en la Universidad Carlos III de Madrid a través de recolectores como e-Ciencia y Recolecta.
- Visibilidad internacional a través de OpenAire y otros agregadores de documentación académica en abierto como Oaister y Base.
- Posición destacada en el Ranking Web de Repositorios.
- Evaluación positiva por REBIUN[2] (cumplimiento de 22 de los 25 criterios).

5. BIBLIOGRAFÍA

DOS EN UNA: se integran las bases de datos de Información de la Actividad Investigadora y el Archivo Abierto. *Digital 3* (28) (febrero 2012), p. 22. Recuperado en: http://hdl.handle.net/10016/14717

FERNÁNDEZ-DEL-PINO, B., RASERO, V., & TOLEDO, G. (2014). *Implantación de ORCID en la Universidad Carlos III de Madrid*. Recuperado de: http://hdl.handle.net/10016/20127

LORENZO, E. (2014, noviembre 14). El soporte de ORCID en Dspace 5 (y superiores) [Mensaje en un blog]. Recuperado de: http://www.arvo.es/dspace/el-soporte-de-orcid-en-dspace-5-y-superiores/

LUYTEN, B. & WOOD, W. (2015). Authority Control of Metadata Values. [Mensaje en una web]. Recuperado de: https://wiki.duraspace.org/display/DSDOC5x/Authority+Control+of+Metadata+Values

ROMÁN, A. (2014, marzo 4). Métodos usados en el Authority Control [Mensaje en un blog]. Recuperado de: http://www.arvo.es/dspace/tag/control-de-Autoridades/

[2] REBIUN. *Evaluación de repositorios institucionales de investigación*. *e-Archivo* fue evaluado en noviembre de 2016. Resultados disponibles en: http://www.rebiun.org/repositorios/Paginas/evaluacion-repositorios.aspx.