INTEROPERABILIDAD ENTRE SISTEMAS DEL INSTITUTO ESPAÑOL DE OCEANOGRAFÍA

Concha Mosquera-de-Arancibia[1], Sergio Nieto-Caramés[2]

[1] Oceanógrafa coordinadora del Repositorio Institucional de Acceso Abierto del Instituto Español de Oceanografía. [2] Ingeniero Informático responsable de Desarrollos Dspace en la empresa Arvo Consultores y Tecnología SL

RESUMEN

Aun cuando habitualmente en las primeras fases del ciclo de implantación de un RI (repositorio institucional), este se considera como un sistema aislado, inmediatamente, aparecen necesidades para que el RI interopere con el resto de sistemas de la organización, pasando a formar parte de lo que podríamos denominar el ecosistema de investigación institucional. Como elementos de esta necesaria interoperabilidad sistémica, destacaríamos, como pieza clave para la simplificación del sistema de trabajo ofrecido a los investigadores de recursos marinos y oceanografía, la integración de e-IEO (Repositorio Institucional Digital de Acceso Abierto del Instituto Español de Oceanografía) con SIPI (Sistema de Seguimiento Integrado de Proyectos de Investigación), que forma parte del Sistema de Gestión de Investigación del IEO. Se presentan los diversos elementos que el Instituto Español de Oceanografía está incorporando a su repositorio para lograr, de forma efectiva, la integración con el resto de sistemas de la organización, entre los que destacamos: 1. Integración de DSpace y SIPI, describiendo los elementos tecnológicos, procedimentales y organizativos necesarios para lograr la interoperabilidad semántica y técnica entre ambos sistemas y evitar a los investigadores el doble archivo. 2. Estandarización de los nombres de autor, mediante la implantación del control de autoridades (authority control) de DSpace, haciendo hincapié en el modelo elegido en e-IEO, en los requerimientos para aplicar esta mejora y en las ventajas obtenidas. 3. Adopción de identificadores ORCID (Open Researcher and Contributor ID) para la identificación y detección de ambigüedades y duplicidades en los nombres de los investigadores, usando un identificador único de autor de amplia difusión y uso. ORCID está conectado a otros sistemas actuales de identificación de autor como Author Resolver, Inspire, IraLIS, Scopus Author Identifier y otros. 4. Inclusión de vocabularios controlados específicos de ciencias marinas (recursos marinos, pesquerías, etc.) como mecanismo de mejora de las capacidades de búsqueda y descubrimiento de los ítems, y para posibilitar la interoperabilidad semántica e incrementar la visibilidad de los ítems en repositorios y recolectores.

Palabras clave: Instituto Español de Oceanografía, repositorio institucional digital, DSpace, interoperabilidad, control de autoridades, identificadores únicos de autor, vocabularios controlados.

ABSTRACT

Interoperability between systems of the Spanish Institute of Oceanography

Even though in the early stages of the cycle of implementation of an IR (institutional repository), this is considered as an isolated system, immediately, are needs that the IR interoperate with other systems in the Organization, becoming part of what we might call the ecosystem of institutional research. As this required systemic interoperability elements, would emphasize, as key for the simplification of the system of work offered to researchers of Oceanography and Marine Resources, the integration of e-IEO (Spanish Institute of Oceanography digital institutional open access repository) with SIPI (database of monitoring integrated research projects), which is part of the system of management of research of the IEO. The various elements that the Spanish Institute of Oceanography is incorporating into your repository to achieve effective integration with other systems of the Organization, among which we highlight are: 1. Integration of DSpace and SIPI, describing the technological, procedural and organizational elements required to achieve semantic and technical interoperability between the two systems and the researchers avoid the double file. 2. Standardization of the names of the author, through the implementation of authority control from DSpace, emphasizing the model chosen in e-IEO, the requirements to implement this improvement and the gains. 3. Adoption of identifiers ORCID (Open Researcher and Contributor ID) for the identification and detection of ambiguities and duplications in the names of the researchers, using a unique identifier of author of wide dissemination and use. ORCID is connected to other current author identification systems as Author Resolver, Inspire, IraLIS, Scopus Author Identifier and others. 4. Inclusion of specific controlled vocabularies of marine sciences (marine resources, fisheries, etc.) as a mechanism for improving capabilities of search and discovery of the items, and to enable semantic interoperability and increase the visibility of items in repositories and harvesters.

Keywords: Spanish Institute of Oceanography, digital institutional repository, DSpace, interoperability, authority control, author identifiers, controlled vocabularies.

INTRODUCCIÓN

En esta comunicación se presentan los sistemas de información incorporados al ecosistema de investigación institucional del IEO, considerado este como el conjunto de elementos de apoyo, soporte y difusión de la actividad investigadora del Instituto, y en el que destacaríamos como elementos principales los siguientes:

- Repositorio Institucional Digital de Acceso Abierto del Instituto Español de Oceanografía, e-IEO.
- Sistema de Seguimiento Integrado de Proyectos de Investigación, SIPI.
- Base de Datos de Autoridades de Autor, BDAA.

La interoperabilidad sistémica que se construye entre estos sistemas, producto de actuaciones en los ejes tecnológicos, procedimentales y organizativos, es vital para lograr una consistencia efectiva entre los datos manejados e intercambiados, con el objetivo final de valorizar la producción investigadora de la Institución y de sus investigadores.

INTEROPERABILIDAD CON EL SISTEMA DE SEGUIMIENTO INTEGRADO DE PROYECTOS DE INVESTIGACIÓN

Desde mediados del año 2014, se han separado funcional y operativamente los sistemas de gestión de resultados de investigación del IEO. Por una parte el sistema de Seguimiento Integrado de Proyectos de Investigación, SIPI, se usa para la gestión interna de los resultados de investigación del IEO, mientras que el repositorio e-IEO, es usado para la ingesta, revisión y exposición de los objetos de investigación.

Esta separación operativa evita la doble descripción de los objetos digitales en SIPI y e-IEO, inasumible por los investigadores, pero en cambio obliga a sincronizar los datos de DSpace en la plataforma SIPI. Esta alineamiento, que básicamente consiste en lograr la interoperabilidad semántica entre ambos sistemas, se logró mediante tres mecanismos principales:

- Alinear los elementos de metadatos descriptivos
- Alinear los valores de metadatos
- Conectar y sincronizar ambos sistemas, e-IEO y SIPI

El primer punto se abordó mediante la inclusión de más de 30 elementos nuevos al esquema Dublin Core Cualificado usado por DSpace (que está derivado del Library Application Profile del DCMI-Libraries Working Group). Se escogió realizar una extensión sobre este esquema frente a la opción de añadir un nuevo esquema específico del proyecto, principalmente por la dificultad de gestionar adecuadamente un nuevo esquema en los procesos de ingesta y exportación masivos.

Este crecimiento de elementos de metadatos implicó a su vez un proceso de adaptación de los formularios de entrada de datos y de visualización de ítems en DSpace, para dar cabida a esos 30 elementos añadidos.

El segundo reto era conseguir adaptar entre ambos sistemas los valores admisibles de los elementos de metadatos utilizados. Entre los elementos tecnológicos que se han incorporado de forma amplia a DSpace tenemos los esquemas semánticos, y bajo este término incluimos vocabularios controlados, tesauros y listados de encabezamiento de materias. Señalar, para dar una idea de la adaptación requerida, que en la versión actual de e-IEO, 17 elementos de metadatos incorporan alguno de estos esquemas semánticos. Éstos son principalmente listas de valores controlados, 11 elementos, pero incluyen también valores controlados de autoridad, 5 elementos, que se detallan en apartado siguiente y un elemento con un vocabulario controlado temático asociado.

Por último, como modelo de conexión y sincronización del e-IEO con el Sistema SIPI, se ha optado por la integración a través ficheros de intercambio, de los contenidos nuevos o actualizados de DSpace. De forma periódica, DSpace pone a disposición de SIPI la información sobre los ítems nuevos o modificados que hayan aparecido en DSpace, en los requeridos formatos de intercambio, con el fin de que SIPI pueda incorpora la información descriptiva asociada a los objetos digitales.

El repositorio e-IEO genera un fichero de texto, en un formato de intercambio definido conjuntamente entre los equipos responsables de ambos sistemas, cada vez que se detecta un cambio en alguno de los objetos almacenados. SIPI, al recibir esos ficheros, los procesa e incorpora así estos cambios a sus procesos.

Para la generación automática de los ficheros de intercambio se ha añadido un proceso a la cola de eventos de DSpace y configurado para generar un evento de un tipo cambio de ítem cada vez que se produce una creación, modificación o eliminación de un metadato existente o de un objeto. Estos eventos se añaden a la cola de eventos que es procesada cada vez que un evento ocurre para evitar sobrecarga en el sistema. Cada evento procesado genera un fichero con un nombre estructurado del tipo "AAMMDDHHMMSSTIPOITEM.txt". Esta nomenclatura simplifica el procesado de los ficheros por SIPI, pues así, cambios sucesivos de un mismo ítem de DSpace se ejecutarán también en el orden cronológico correcto en SIPI y evitaremos la posible inconsistencia de los metadatos.

CONTROL DE AUTORIDADES

El e-IEO usa el modelo de autoridades de DSpace (plugin de DSpace, Authority Control) para la validación interna de sus autores, implementado este sobre el metadato dc.contributor.author, principalmente, aunque se extiende al resto de elementos de la tipología dc.contributor.x.

Este metadato está conectado con uno los registros de autores del IEO, en forma de Base de Datos de Autoridades de Autor, BDAA, que contiene los nombres de los autores de la institución y los códigos de autoridades, entre otros datos. En el caso del e-IEO los registros de autoridad, contienen la forma autorizada del nombre del autor, establecida por la normativa DRIVER (DRIVER 2.0., 2008) como forma preferida, así, por .ejemplo, el autor José Francisco Domínguez Yanes, se describe en la BDAA, y por ende en el repositorio e-IEO, según su nombre normalizado DRIVER, Domínguez-Yanes, J.F. (José Francisco).

Los objetivos del modelo de autoridades implantado en el e-IEO son los siguientes:

- Dar consistencia e integridad a los metadatos, ayudando en la corrección de los correspondientes valores. En el repositorio e-IEO se ha conseguido conectando las interfaces de archivo y edición de DSpace con una base de datos de autores del IEO con el fin de chequear los valores introducidos contra los registros de autoridad y poderles asignar una clave de autoridad única.
- Conseguir mejorar la precisión en la recuperación de la información, puesto que el mejor método, simple y positivo, de determinar si dos valores son idénticos, es comparando las claves de autoridad, ya que comparar valores textuales proporciona falsos positivos (demasiados García, M.) o falsos negativos (¿García, M. vs. García, Manuel?)
- Facilitar el intercambio de información bibliográfica con el SIPI, puesto que en la transferencia de información del e-IEO a SIPI se envían los valores de clave de autoridad, reconocibles y comunes entre ambos sistemas, en vez de los valores textuales de los nombres de autor.

En el e-IEO se usa, además, el modelo de *valores de confianza (confidence values)* de DSpace para mejorar la operativa del sistema de validación de autores. El valor de confianza se asigna adicionalmente al valor de clave de autoridad y se expresa como un valor simbólico dentro del rango *aceptado, incierto, ambiguo, no encontrado* y *sin validar*. Los valores de confianza aplicados en el e-IEO son los siguientes:

- *Aceptado:* el autor ha sido validado por un usuario (con permisos de envío, edición o revisión)
- *Incierto:* el autor ha sido validado por los procesos automatizados de validación de autores. En los procesos del repositorio, este valor de confianza tiene el mismo uso que el valor de aceptado.
- *Ambiguo:* se han encontrado varias posibles coincidencias en la BDAA, por lo que el usuario ha de desambiguar el valor en un proceso de edición manual.
- *No encontrado:* el autor no pertenece a la BDAA.
- *Sin validar*: el autor está pendiente de validar.

El modelo de Authority Control se aplica en el e-IEO por medio de varios procesos o en diversos momentos del ciclo de vida de un ítem:

- Proceso de depósito de ítems. Mediante la validación manual al insertar un nuevo ítem en el repositorio. Consiste en validar los autores mediante una pantalla de validación en el formulario de envío. Si se valida un autor contra la BDAA, a dicho autor se le normaliza el nombre al valor almacenado en la BDAA, se le asigna una clave de autoridad y se le asigna un valor de confianza de *aceptado*.
- Proceso de edición/revisión de ítems. Para los ítems archivados que requieran la modificación de los metadatos controlados por autoridad, dc.contributor.author, se puede aplicar el mismo proceso que en el depósito de ítems. Mediante una ventana el editor/revisor puede visualizar todos los valores de autor que coincidan parcial o totalmente con el nombre almacenado. Al seleccionar uno, se efectúa la validación del autor, con las acciones anteriormente citadas: normalización de nombre, asignación de clave de autoridad y asignación de valor de confianza.
- Tareas de curación. Debido al alto número de ítems que requieren validación de autores, ya que existen procesos de carga masiva de objetos provenientes de sistemas externos, en el e-IEO se implantó un tarea de curación que efectúa la validación automática de autores contra la BDAA. Este proceso se explica en los párrafos siguientes.

Para facilitar la validación de autores, se han implementado procesos de curación, procesos que pueden ejecutarse para comunidades, colecciones o ítems aislados. Estos procesos validan el metadato dc.contributor.author contra la BDAA. De entre los métodos disponibles en el Authority Control de DSpace se ha escogido el getBestMatch (lograr la mejor coincidencia) que confronta el valor del metadato 'nombre y apellidos del autor' con la BDAA, con los siguientes resultados:

- Si hay una única coincidencia con la BDAA, se asigna la clave de autoridad correspondiente, se normaliza el nombre y se asigna un valor de confianza de *incierto* (pue
- sto que el proceso se ha realizado sin intervención humana, no se asigna al valor *aceptado*, aunque son valores de confianza funcionalmente equivalentes)
- Si hay más de una coincidencia, no se asigna clave de autoridad ni se normaliza el nombre, pero se le asigna un valor de confianza de *ambiguo,* indicando que un usuario deberá de desambiguar manualmente el nombre de autor.
- Si no hay coincidencias, se asigna un valor de confianza de *no encontrado*.

IDENTIFICADORES ÚNICOS DE AUTOR

ORCID es un registro abierto de identificadores de investigadores y autores, con enlace a sus publicaciones, que nace como servicio en el año 2012. Pretende convertirse en un estándar para la identificación única y persistente de autores.

Aunque inicialmente su crecimiento se sustentó en el registro individual por los propios autores (en febrero de 2013 tenía 60.000 registros de autor), en la actualidad sus cifras parecen derivar del registro de Instituciones, Editores, Proveedores de servicios de Información, Integradores de software, etc. (110 miembros en julio 2014) y la consiguiente afiliación automática de sus autores (en julio de 2014, 800.000 registros).

Aunque el IEO participa en e-Ciencia (proyecto enmarcado en el convenio de cooperación interbibliotecaria entre la Comunidad de Madrid y el Consorcio Madroño para crear una plataforma digital de acceso abierto a la producción científica de la Comunidad de Madrid) y Madroño, Consorcio de Universidades de la Comunidad de Madrid para la Cooperación Bibliotecaria, adherido a ORCID, aún no ha efectuado un proceso de afiliación masiva de sus aproximadamente 700 investigadores y autores. No obstante, un número indeterminado de estos ha realizado su registro en ORCID y IEO ha considerado necesario empezar a incorporar funcionalidades de integración con la plataforma orcid.org para la validación internacional de sus autores.

En esta primera fase, el identificador ORCID, para los autores que dispongan del mismo y que lo notifiquen a los responsables del repositorio, se ha incluido en la BDAA, lo que posibilita que dichos identificadores se puedan usar para ofrecer servicios de conexión entre sistemas. Esto se consigue mediante la creación de relaciones de enlace entre los valores de autoridad locales, derivados de la BDAA y los valores de autoridad externos de ORCID. El IEO considera que al exponer sus autores validados a los usuarios finales, estas relaciones entre sistemas adquieren cada vez mayor importancia, "*permitiendo a los repositorios ofrecer servicios de impacto a pesar de los pocos datos disponibles en los repositorios institucionales comparados son los sistemas globales"* (Tarver *et al.*, 2014).

Señalar que la inclusión de identificadores de autor externos adicionales posibilitará una nueva generación de servicios orientados al investigador, identificándose diversas áreas de actuación (ORCID, 2014), entre las que destacaríamos de interés para el e-IEO las siguientes:

- Enlace desde el perfil del autor en e-iEO al registro ORCID en orcid.org
- Simplificación del autoarchivo, conectando este proceso con los registros de publicaciones de un autor a través de su identificador ORCID, evitando la doble introducción de datos y mejorando la calidad de los registros.
- Sincronización de publicaciones entre e-IEo y los registros orcid.org
- Uso del identificador ORCID como clave de autoridad de autor, cuando la adopción de este identificador sea generalizada por la comunidad investigadora.

VOCABULARIOS CONTROLADOS DE MATERIAS

Cuando los repositorios temáticos como el e-IEO se plantean ofrecer servicios y ser visibles por una audiencia más allá de sus fronteras institucionales, es esencial la descripción de contenidos y su indexado por medio de metadatos estandarizados que sean relevantes, tanto en el proceso de submisión como en el descubrimiento de recursos (Subirats *et al*., 2012) y por esto es especialmente relevante el uso de tesauros temáticos, considerados éstos como un conjunto de "términos" empleados para representar los conceptos, temas o contenidos de los documentos de un disciplina o temática específica.

La incorporación de metadatos estandarizados por medio de este tipo de vocabularios controlados tiene las ventajas de a) mejorar las capacidades de búsqueda y descubrimiento de los ítems; b) mejorar o posibilitar la interoperabilidad (semántica) con repositorios o recolectadores temáticos y c) incrementar la visibilidad de los ítems en repositorios y recolectores, por ejemplo en OCLC).

Los vocabularios controlados en dominios temáticos específicos tienen un amplio potencial cuando se usan en la recuperación de información, principalmente por su posibilidad de incorporar términos relacionados usados en el indexado, así como términos multi-lenguaje, (Borst, 2012) aspectos éstos que se han valorado en el repositorio e-EIO.

En línea con otros plataformas de repositorio de ciencias marinas, cuya referencia principal es AgriOcean DSpace, desarrollo conjunto de la FAO, agencia de las Naciones Unidas y de Unesco-IOC/IODE, se ha incorporado al e-IEO un subconjunto del vocabulario especializado AGROVOC. Igualmente se está evaluando la incorporación al repositorio en una fase posterior del vocabulario especializado Aquatic Sciences and Fisheries Abstracts (ASFA).

AGROVOC cubre todas las áreas de interés de la FAO, incluyendo alimentación, nutrición, agricultura, pesca, etc, por lo que la implementación para el e-IEO es un subconjunto del vocabulario, incorporando las subclasificaciones de recursos marinos, acuáticos y pesqueros, y que en forma de esquema SKOS-XL está disponible para organizaciones de todo el mundo.

Al ser un vocabulario que en su versión amplia maneja 32.000 conceptos en 21 idiomas, su tratamiento por DSpace tiene que realizarse, por motivos obvios de rendimiento, mediante la incorporación de un servidor de vocabularios específico, ASKOSI, mejorando significativamente la usabilidad de la solución construida.

REFERENCIAS BIBLIOGRÁFICAS

Borst, T. (2012). Usage and Impact of Controlled Vocabularies in a Subject Repository for Indexing and Retrieval. In *Liber Quarterly* Volume 21 Issue 3/4 2012 pp.: 445- 453.

DRIVER 2.0. (2008). Directrices para proveedores de contenido - Exposición de recursos textuales con el protocolo OAI-PMH. Digital Repository Infraestructure Vision for European Research.
http://www.driver-support.eu/documents/DRIVER_2_0_Guidelines_Spanish.pdf

ORCID (2014). ORCID Member Integration Guide.
http://orcid.org/organizations/integrators. Accessed July 30, 2014.

Subirats, I.; Malapela, T.; Dister, S.; Zeng, M.; Goovaerts, M.; Pesce, V. y Keizer, J. (2012). Reorienting open repositories to the challenges of the Semantic Web: Experiences from FAO's contribution to the resource processing and discovery cycle in repositories in the agricultural domain. In *Metadata and Semantics Research,* pp.: 158-167. Springer Berlin Heidelberg.

Tarver, Hannah; Waugh, Laura; Phillips, Mark Edward y Hicks, William. (2014). *Implementing Name Authority Control into Institutional Repositories: A Staged Approach*. UNT Digital Library.
http://digital.library.unt.edu/ark:/67531/metadc172365/. Accessed July 30, 2014.