Reporte Técnico

Reestructuración y normalización de palabras claves en SEDICI

de Albuquerque, Pablo César
Centro de Servicios en Gestión de Información (CESGI), Comisión de Investigaciones Científicas de la Provincia de Buenos Aires
PREBI-SEDICI, Universidad Nacional de La Plata
pablo@sedici.unlp.edu.ar
https://orcid.org/0000-0001-5277-1665

La Plata, 15 de Junio de 2020

Este informe interno es el resultado del trabajo realizado en el repositorio institucional de la Universidad Nacional de La Plata, SEDICI (http://sedici.unlp.edu.ar), con el objetivo de describir el proceso de reestructuración del esquema de metadatos cuyos valores provienen de vocabularios controlados preexistentes y son utilizados como palabras claves.

Introducción

El repositorio institucional de la Universidad Nacional de La Plata combina esquemas de metadatos estandarizados con uno propio para describir los recursos almacenados. Algunos de estos metadatos obtienen sus valores a partir de entidades normalizadas llamadas autoridades, que conforman vocabularios controlados, como tesauros y taxonomías . Estos pueden ser elaborados por otras instituciones, organizaciones especializadas en una temática particular o de manera interna ajustándose a las necesidades propias de un repositorio, con el fin de asistir al usuario durante la carga de contenido, normalizando la información de los recursos, mejorando la interoperabilidad de un repositorio con otros sistemas, y organizando los recursos mejorando procesos como por ejemplo el de recuperación de contenido (*Name Authority Control in Institutional Repositories and Its Relationships to Metadata Quality*, 2011). La forma en que se controla un valor almacenado es relacionándolo con el identificador de la una autoridad (*Authority Control of Metadata Values - DSpace - LYRASIS Wiki*, 2020).

Previo a las modificaciones detalladas en este informe, el esquema utilizado en el repositorio SEDICI disponía de varios metadatos para almacenar valores que se utilizaban como palabras claves de un recurso. Una “palabra clave” es un concepto que puede estar formado por una o más palabras y se relaciona a un recurso con el fin facilitar el descubrimiento del mismo por usuarios o programas (*Guides: Health Sciences Research Basics: Subject Headings/ Keywords*, 2020). En SEDICI estas palabras claves pueden estar controladas por una autoridad o no.

El esquema de metadatos mencionado anteriormente dedicaba un metadato para cada fuente de valores disponible, haciendo que existan palabras claves dispersas en varios metadatos y obligaba a que el usuario que realiza la carga de recursos deba tener un conocimiento previo para saber qué valor se podía recuperar de cada vocabulario.

Los metadatos en cuestión eran:

Metadato	**Descripción**
sedici.subject.lcsh	Materia asociada al documento, extraída del vocabulario del Library of Congress Subject Headings
sedici.subject.decs	Descriptores extraídos del tesauro DECS
sedici.subject.eurovoc	Descriptores extraídos del tesauro Eurovoc
sedici.subject.descriptores	Descriptores extraídos de los tesauros SeDiCI y Unesco
sedici.subject.other	Descriptores libres
sedici.subject.keyword	Palabras clave extraídas del documento
sedici.subject.acmcss98	Descriptores extraídos del Sistema de clasificación ACM CSS 98

Cada uno de estos metadatos se corresponde con un campo en el formulario de carga, este diseño conlleva a que los formularios sean complejos y traigan ciertas desventajas como:

1. Se dificulta su interpretación y se disminuyen los tiempos de carga.
2. El usuario muchas veces no completa los valores para un determinado metadato, o solo lo hace para uno y no para el resto.
3. Algunos campos pueden quedar vacíos ya que se completan con valores obtenidos de otros vocabularios.

Este diseño hace que en caso de que se quiera conocer las palabras claves para un recurso, se deba consultar a varios metadatos, haciendo más compleja la consulta y teniendo que diferenciar valores que se repitan.

Otro inconveniente que presentaba SEDICI es la falta de mantenimiento de los vocabularios consultados, esto que algunos valores ingresados no sean asociados con palabras claves que existen en nuevas versiones, lo que obliga a quien carga, a ingresar valores que sin controlar aumentando la posibilidad de que estos tengan errores sintácticos o no puedan ser referenciados como una entidad válida por un tercero.

En términos de base de datos, el hecho de disponer de vocabularios compuestos por términos que pueden estar desactualizados, sumado a que un término puede repetirse en varios vocabularios lleva a un mantenimiento innecesario de los mismos, dificultando esta tarea.

Más allá de estas cuestiones que afectan el uso interno de las palabras claves en el repositorio, el hecho de asociar un recurso con palabras claves ofrece un mecanismo sencillo y eficiente para agrupar los recursos relacionados. El objetivo de este trabajo es disponer de un conjunto de palabras claves normalizadas, interoperables y vigentes (actualizadas), para vincular conceptos (representados en una o varias palabras claves) con los recursos almacenados en el repositorio, de esta forma mientras más representativa sea la palabra clave, mejores agrupaciones se podrán realizar aumentando el número de resultados útiles devueltos por una búsqueda determinada

Trabajo realizado

Como primera medida, se buscaron vocabularios controlados que pudiesen relacionarse con los valores sin controlar en los metadatos mencionados en el punto anterior. Para la elección de los vocabularios se realizó un análisis de cobertura de los vocabularios utilizados por otras instituciones.
Luego de las consideraciones anteriores, se propuso migrar todos los valores de los metadatos destinados a palabras claves a un único metadato "*dc.subject*".

Esta tarea implicó:

- Normalización de metadatos.
- Análisis de cobertura de diversos vocabularios controlados con los valores almacenados en SEDICI que no tienen un clave de autoridad asociada.
- Identificación de vocabularios controlados que se utilizarán a partir de los resultados a anteriores.
- Importación de los nuevos vocabularios al sistema de de gestión de vocabularios controlados para SEDICI
- Migración de valores existentes a un único metadato

Esta migración trajo las siguientes ventajas:

1. Aumento en la calidad de los metadatos con valores actualizados.
2. Mejora en la navegación y búsqueda de contenidos
3. Formularios de carga más sencillos
4. Esquema de metadatos más sencillo y fácil de mantener

Normalización de metadatos

Previo a la realización de un análisis de cobertura fue necesario aplicar un pre procesamiento a los datos almacenados, ya que para un mismo concepto existían variantes sintácticas debido a errores de carga, usos de mayúsculas y a que a veces se ingresaban valores en singular y otros en plural. Por ejemplo existían 167 recursos en el repositorio con la palabra clave "jóvenes" que no se relacionaban con ninguna autoridad, cuando en el tesauro de la UNESCO existe el término "joven".
Para esto se agruparon los valores y se contabilizó la cantidad de veces que se repetían los términos en la base de datos. Los términos con faltas de ortografías fueron fáciles de corregir ya que eran aquellos que no estaban controlados y tenían menos repeticiones.

Análisis de cobertura

Para determinar qué vocabularios podían ser utilizados se trabajó en conjunto con el personal de carga de SEDICI. Se decidió analizar los vocabularios de UNESCO, AGROVOC, DECS, GeoNames y ACM, debido a que son mantenidos por organizaciones de renombre y son ampliamente utilizados.

Tesauro de la UNESCO

Vocabulario controlado disponible en varios idiomas. Su versión en español está compuesto por 4471 términos destinados al análisis temático y la búsqueda de documentos y publicaciones en los campos de la educación, cultura, ciencias naturales, ciencias sociales y humanas, comunicación e información. Este vocabularios se encuentra en desarrollo continuo y su terminología multidisciplinaria refleja la evolución de los programas y actividades de la UNESCO (*Tesauro de la UNESCO*, 2020).

AGROVOC

AGROVOC es un vocabulario controlado publicado por la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) que abarca todos los ámbitos relacionados a la alimentación, la nutrición, la agricultura, la pesca, las ciencias forestales y el medio ambiente.
AGROVOC consta de más de 35,000 conceptos, disponibles en varios idiomas (*AGROVOC tesauro multilingüe de agricultura | Agricultural Information Management Standards (AIMS)*, 2020).

DeCS

El vocabulario DeCS (Descriptores en Ciencias de la Salud) es utilizado para la indización de artículos de revistas científicas, libros, anales de congresos, informes técnicos, y otros tipos de materiales, así como para ser usado en la búsqueda y recuperación de asuntos de la literatura científica en las fuentes de información disponibles en la Biblioteca Virtual en Salud (BVS) como LILACS, MEDLINE y otras. (*DeCS - Descriptores en Ciencias de la Salud*, 2020)

Fue desarrollado a partir del MeSH - Medical Subject Headings de la U.S. National Library of Medicine (NLM) con el objetivo de permitir el uso de terminología común para búsqueda en múltiples idiomas, proporcionando un medio consistente y único para la recuperación de la información.

El DeCS es un vocabulario dinámico totalizando 34.118 descriptores y calificadores, siendo de estos 29.716 del MeSH y 4.402 exclusivamente del DeCS. Este vocabulario es actualizado anualmente por MeSH

GeoNames

GeoNames es una base de datos geográfica con un fuerte grado de adopción a la web semántica compuesta por más de 10 millones de nombres geográficos. Para cada lugar se poseen datos como la latitud, longitud, altitud, población entre otros. (*About GeoNames*, 2020)

ACM CCS 2012

El Sistema de Clasificación de Computación ACM (ACM CCS) del año 2012 ha sido desarrollado como un reemplazo de la versión de 1998 y ha sido adoptado como el sistema de clasificación estándar para el campo de la computación. Se está integrando en las capacidades de búsqueda y en las visualizaciones de temas visuales de la Biblioteca Digital ACM. (*The 2012 ACM Computing Classification System*, 2020)

Resultado de análisis de cobertura

Para buscar relaciones entre los datos en el repositorio con otros vocabularios se tuvo en cuenta que pueden existir variantes sintácticas, para eso todos los términos fueron llevados a minúscula y se quitaron los acentos. También se aplicaron varias funciones de comparación de texto como Levenshtein y Soundex (*PostgreSQL: Documentation: 12: F.15. fuzzystrmatch*, 2020), aunque estos no dieron buenos resultados ya que el primero tardaba demasiado y no devolvía buenos resultados mientras que el segundo funciona bien solo para palabras en inglés y las palabras a analizar estaban mayormente en español.
Una vez que se encontraron los términos controlados que coincidan exactamente con los valores almacenados en SEDICI restaba ver los términos que difieren dado su número (singular o plural). Para esto se buscaron los valores utilizados como palabras claves, sin controlar, que terminen con ‘S’ o con ‘ES’, y se los comparó con los términos controlados obviando estos últimos caracteres.
Una vez completadas estas pruebas se continuó con la importación de los vocabularios DeCS 2018, ACM CCS 2012 y UNESCO. La decisión de incorporar UNESCO es debido a su moderado tamaño (4471 términos) de los cuales un 45.18 % coincide con los valores en los metadatos a migrar existentes. DeCS se continuaría usando dado que la administración del repositorio ya tenía experiencia en su uso, es demasiado específico y a pesar de ser muy grande (31884 términos en español) existe un 18.13 % de coincidencias con los valores en sedici.subject

Tanto Agrovoc como GeoNames no fueron incorporados. El motivo por el cual se descartó el uso de Agrovoc fue debido a que a pesar de su gran tamaño (34060 términos en español) solo aproximadamente un 11 % tenía correspondencia con los valores de SEDICI. De ese 11 % más de la mitad se superpone con términos en DeCS y UNESCO, por lo que incorporar un vocabulario nuevo tan grande, para solo incorporar cerca de un 5 % de términos nuevos, significaba demasiado trabajo y posterior mantenimiento. Se descartó el uso de GeoNames, al menos por ahora, ya que como en esta etapa la comparación que se realizó entre los metadatos y los términos fue sintáctica, se iban a generar muchos problemas con los nombres de personas que coinciden con lugares (por ejemplo “San Martín”).

Importación de vocabularios

Los vocabularios fueron incorporados como taxonomías en un sistema basado en Drupal que expone sus recursos al repositorio a través de SPARQL (de Albuquerque, 2018). Esto trae como ventaja poder recorrer y navegar el vocabulario como un árbol, pero trae como desventaja la duplicación de algunos nodos, ya que si un término tiene más de un padre necesariamente este aparecerá en dos ramas distintas. Como no se utiliza la URI del recurso

en Drupal como clave de la autoridad, sino el identificador del término propio del vocabulario, la duplicación de nodos no resulta un inconveniente.

Migración de valores existentes a un metadatos

Se movieron los valores en los metadatos bajo sedici.subject a dc.subject y se realizó la configuración en SEDICI para poder consumir información de estos nuevos vocabularios. Hasta ese momento, cuando un usuario realizaba la búsqueda de una palabra clave, el software DSpace generaba una consulta SPARQL de tipo CONSTRUCT y Drupal retornaba un conjunto de grafos RDF. Cada uno de estos grafos RDF estaba compuesto por nodos que contenían la autoridad (palabra clave e identificador) que coincidía con la búsqueda junto con otros nodos que se relacionaban con esta respuesta, por ejemplo la autoridad padre o información sobre el vocabulario al que pertenecía la autoridad. Esto le permitía a DSpace procesar ese grafo y mostrarle al usuario información contextual de esa respuesta, que le permitiese al usuario entender a que hace referencia ese concepto de forma más precisa. En este punto, solo existía en este sistema de gestión el vocabulario DeCS (73292 nodos, incluyendo los duplicados). Cuando un usuario buscaba una palabra clave lo hacía solamente dentro de este vocabulario, logrando tiempos de respuesta aceptables. Sin embargo una vez que se sumaron los vocabularios, AGROVOC, ACM CCS 2012 y UNESCO a Drupal, se observó una caída en la performance de los tiempos de respuesta, debido al gran volumen de datos en el que había que buscar sumado al procesamiento para el armado de las respuestas, llegando a durar estas entre 1,5 y 4 segundos.
Se realizó una consulta con los usuarios administradores de SEDICI, y se observó que no era necesario realizar tanto procesamiento por parte del repositorio, ya que como ahora todas las palabras claves se recuperan desde un solo campo no les resultaba necesario el nombre del padre de un término ni el nombre del vocabulario que ofrecía la autoridad. Dado esto último, se decidió recuperar los términos a través de una consulta SPARQL de tipo SELECT, más eficiente y sencilla, que solo devuelve el texto y el identificador de la autoridad. De esta forma se evitó que el software DSpace realice un procesamiento innecesario más adelante.

Conclusión

Se realizaron modificaciones en el esquema de metadatos de SEDICI, pasando de un modelo donde las palabras claves se encontraban en varios metadatos a un único metadato, que almacena valores que pueden estar controlado por autoridades provenientes de vocabularios externos, elegidos en base a las necesidades propias del repositorio. Los valores que son muy utilizados en el repositorio, pero no se les ha encontrado una autoridad que los controle, son candidatos a formar parte de un vocabulario propio a ser implementado más adelante.
Este enfoque permite actualizar o agregar nuevos vocabularios al sistema de gestión de autoridades para ser utilizados como palabras claves por el repositorio de manera transparente, ya que este no debe necesariamente conocer el origen del término controlado sino que le basta con saber que ese valor es correcto.
El hecho de disponer de un conjunto de palabras claves normalizadas, en forma de autoridades, mejora la navegación y la búsqueda de recursos asociados a estos términos y posibilita interoperar con otros sistemas que hagan referencia a estas mismas autoridades a través de su identificador.

Bibliografía

About GeoNames. (2020). https://www.geonames.org/about.html

AGROVOC tesauro multilingüe de agricultura | Agricultural Information Management Standards (AIMS). (2020). http://aims.fao.org/es/agrovoc

Authority Control of Metadata Values—DSpace—LYRASIS Wiki. (2020). https://wiki.lyrasis.org/display/DSPACE/Authority+Control+of+Metadata+Values

de Albuquerque, P. C. (2018). *Soporte de vocabularios controlados y autoridades en repositorios digitales* [Tesis, Universidad Nacional de La Plata]. http://sedici.unlp.edu.ar/handle/10915/69754

DeCS - Descriptores en Ciencias de la Salud. (2020). http://decs.bvs.br/E/homepagee.htm

Guides: Health Sciences Research Basics: Subject headings/ Keywords. (2020). https://libraryguides.mcgill.ca/healthscibasics/headings-keywords

Name Authority Control in Institutional Repositories and Its Relationships to Metadata Quality. (2011).

Tesauro de la UNESCO. (2020). http://vocabularies.unesco.org/thesaurus

The 2012 ACM Computing Classification System. (2020). https://www.acm.org/publications/class-2012