Soporte de vocabularios controlados y autoridades en repositorios digitales

Capítulo 1 - Introducción

Repositorios Institucionales Digitales

Un repositorio institucional es una infraestructura web cuya finalidad es la de almacenar, organizar, preservar y dar acceso público a la producción digital generada en un institución y que es capaz de interoperar con otros repositorios similares. Para cumplir con estos objetivos es necesario poder describir la producción de la institución como un recurso, es por eso que se utilizan metadatos.

Los metadatos son datos estructurados que sirven para identificar, describir, localizar y facilitar la obtención, el uso y la administración de un recurso. Un repositorio institucional utiliza uno o más esquemas de metadatos, pudiendo ser estos estandarizados o definidos internamente, dependiendo de las necesidades de los mismos y su organización.

La información que describe un recurso puede ser diversa y en algunos casos depende del metadato en cuestión. Por ejemplo, el título de un recurso puede ser representado con una cadena de caracteres cualquiera, no así una licencia de uso donde la elección del valor se ve restringida a las licencias existentes. El uso de vocabularios controlados es una práctica que permite normalizar la información de los recursos y mejorar la interoperabilidad de un repositorio con otros sistemas, proporcionando una forma de organizar el conocimiento mejorando procesos como por ejemplo el de recuperación de contenido.

SEDICI

SEDICI (Servicio de Difusión de la Creación Intelectual) es el repositorio institucional central de la Universidad Nacional de La Plata creado en el año 2003. Actualmente aloja tesis de grado, tesis de posgrado, artículos, revistas producidas por las distintas unidades académicas de la UNLP así como también otros tipos de documentos como artículos, presentaciones en congresos, informes técnicos, proyectos de investigación, ordenanzas y reglamentaciones, entre otros.
Este repositorio cuenta con las siguientes vías de depósito:

- Autoarchivo: Realizado por un autor registrado en el repositorio. Para esto el usuario carga su obra a través de una serie de pasos definidos que incluyen la descripción (tipología, título, etc), adjunción de archivos y selección de licencia de uso.
- Depósito delegado: Se hace entrega de la obra junto con la licencia de depósito siendo la misma cargada en el repositorio por personal de SEDICI.
- Ocasionalmente el repositorio asigna un permiso especial a un representante de otra institución que asume el compromiso de carga de los materiales del mismo.

- A través de un sistema mediado: donde administradores del repositorio agregan contenidos de determinados lugares de la web.

En todos los casos personal del repositorio realiza tareas de revisión, normalización de metadatos para luego publicar o vetar la obra. Es posible aplicar un bloqueo temporal, denominado embargo, sobre los recursos publicados. Durante el transcurso del embargo el contenido de una obra se mantiene oculto e inaccesible.

Motivación

El origen de la información que describe a los recursos depositados en un repositorio institucional es diverso y es por este motivo que estos sistemas deben contemplar el uso de vocabularios controlados para normalizar la información de los recursos, pudiendo mejorar aspectos como la carga de metadatos, la interoperabilidad entre sistemas y la búsqueda y difusión de contenido.

Un vocabulario controlado es una conjunto organizado de palabras utilizadas para indexar, describir y recuperar contenido. Su uso en los repositorios digitales institucionales favorecen múltiples procesos ya que esta práctica define un conjunto limitado de términos para referirse a un único concepto de manera consistente, por ejemplo una persona o un lugar. Esta práctica no solo trae beneficios dentro de la institución, sino que el hecho de poder referenciar unívocamente conceptos ya definidos favorece aspectos como la interoperabilidad entre otros sistemas y el repositorio.

SEDICI se encuentra desarrollado sobre DSpace, un software de código abierto para repositorios digitales que provee herramientas para la gestión, acceso y preservación de documentos digitales. Este software tiene la posibilidad de gestionar internamente un conjunto limitado de vocabularios controlados pero con ciertas limitaciones sobre todo al momento de agregar nuevos vocabularios. Sin embargo DSpace posee una arquitectura de plug-ins que permite implementar un conector cuya finalidad sea la de integrar vocabularios definidos en un sistema de gestión externo y relacionarlos estos con los metadatos del repositorio afectando todas las fases de la gestión de metadatos, brindando las siguientes ventajas:

1. Proveer una forma simple de probar si dos valores son idénticos comparando el identificador de los vocabularios controlados.
2. Asistir al usuario para completar de manera correcta un conjunto de metadatos
3. Mejorar la interoperabilidad del repositorio facilitando el intercambio de información bibliográfica.

Objetivo

El objetivo principal de la tesina es diseñar e implementar un sistema de gestión de autoridades y vocabularios controlados en general para repositorios digitales.

Objetivos específicos:

- Permitir la reutilización de vocabularios controlados estructurados preexistentes como ser tesauros, sistemas de clasificación, entre otros.
- Implementar un sistema genérico para gestión de vocabularios controlados en el ámbito del repositorio SEDICI que sea independiente del software de repositorio.
- Migrar los datos existentes en sistemas previos de SEDICI para que se adapten a la nueva infraestructura.
- Implementar un conector para DSpace[1] que se comunique con el nuevo sistema usando protocolos y formatos estándares.
- Definir un mecanismo replicable para aplicar el módulo de vocabularios controlados en otros repositorios digitales.

Capítulo 2 - Vocabularios controlado

Introducción

Un vocabulario controlado es un conjunto limitado de términos y reglas que forman parte de una estructura con fines de indexación y recuperación de información (COAR, «FAQs for Controlled Vocabularies») que puede incluir términos preferidos y variantes (Bernal, «Uso de Vocabularios Controlados en Repositorios. La experiencia de DIGITAL.CSIC».

La creación de los vocabularios controlados tiene como origen la necesidad de unificar los encabezamientos o puntos de acceso de un catálogo bibliográfico conformado por ficheros de autoridad. Muchas de las bibliotecas más antiguas gestionaban ficheros de autoridad en paralelo al catálogo bibliográfico para uso exclusivo de los bibliotecarios. En estos registros no existía uniformidad en la estructura, puntuación o los signos empleados, pero si coincidían una serie de elementos, como las relaciones de un término con sus términos aceptados y alternativos o entre términos posteriores y anteriores, así como algunas notas con escasa normalización.

El uso de vocabularios controlados en repositorios institucionales digitales resulta beneficioso en el proceso de indexación ya que los repositorios u otros sistemas utilizan el mismo término para referirse al mismo concepto (por ejemplo, persona, lugar o cosa) de una manera consistente. Este mecanismo ayuda la búsqueda y descubrimiento de contenidos.

Los vocabularios controlados asisten a la usuarios no solo en búsqueda específicas de contenido sino también durante la navegación de un sitio a través del uso de filtros generados a partir de subconjuntos o facets. De hecho, la función más útil de los vocabularios controlados es reunir términos variantes y sinónimos de conceptos en una misma entidad. Por lo tanto, la asociación de los diversos sinónimos en un término controlado aumenta el número de accesos útiles devueltos por la búsqueda.

Existen varios tipos de vocabularios controlados, pero en este trabajo solo se hará énfasis en las taxonomías, los catálogos de autoridades, las listas de encabezamientos de materia, los tesauros y las ontologías, ya que estos son los más relevantes para un repositorio.

Tipos de vocabularios controlados

A. Encabezamientos de materia

Los encabezamientos de materia son listas alfabéticas formadas por encabezamientos y subencabezamientos que sirven para designar la temática a un conjunto de documentos

(Biblioteca Nacional de España, «5.6. Encabezamientos de materia»). Los encabezamientos están formados por una o varias palabras que representan conceptos, tratando de condensar el tema sobre el que trata el documento.

Las funciones de las listas de encabezamientos son:

- Describir el contenido de los documentos indexados.
- Agrupar todos los documentos de temáticas afines.
- Evitar la ambigüedad expresando cada concepto en un término.
- Permitir la recuperación de los documentos por el campo materia.
- Relacionar documentos a partir de las relaciones semánticas entre términos.

Ejemplos

LCSH

La Biblioteca del Congreso de los Estados Unidos dispone de un vocabulario controlado de encabezados, el Library of Congress Subject Headings (LCSH), utilizado para catalogar los materiales conservados. El mismo puede ser consultado a través de una interfaz web o puede ser descargado en formatos RDF/XML, N-Triples o JSON (The Library of Congress, «LC Linked Data Service»).

LEMBP

La Lista de Encabezamientos de Materia para las Bibliotecas Públicas (LEMBP) es una lista de materias adaptada a los principios de los datos abiertos vinculados (Linked Open Data).
La LEMBP se encuentra representada en formato SKOS y ha sido ampliada con encabezamientos de otros registros de autoridad (como el de la Biblioteca Nacional española) y vinculada con otras listas internacionales como LCSH y RAMEAU (lista de encabezamientos utilizada en la red de bibliotecas públicas francesas y en la Biblioteca Nacional Francesa).
Los encabezamientos pueden ser recuperados mediante un buscador simple, un buscador avanzado o un endpoint SPARQL, además se pueden descargar ficheros con las distintas listas completas, tanto en formato MARC21 como SKOS (RDF/XML).

MESH

MeSH (Medical Subject Headings) es un tesauro cuyo objetivo principal es proporcionar una terminología jerárquicamente organizada para indexar y catalogar información biomédica, de artículos en MEDLINE/PUBmed y otras bases de datos de la National Library of Medicine (NLM). (MESH, «Medical Subject Headings - Home Page».)
El tesauro MeSH puede ser consultado a través de una interfaz web, puede ser descargado en archivos en formato ASCII, XML, MARC y RDF.

MeSH puede ser consultado a través de un endpoint SPARQL (API) (MeSH, «MeSH Linked Data») y una interfaz RESTful.

B. Catálogo de autoridades

Se define como autoridad a un conjunto de entidades normalizadas, compuestas por un registro y una clave. El registro de autoridad contiene información asociada con una entidad mientras que la clave de autoridad es un código que identifica unívocamente a la misma. Un catálogo de autoridades es un vocabulario controlado constituido por un conjunto de registros normalizados de autoridad (Biblioteca Universidad De Salamanca, «Catálogo “Biblioteca Universidad De Salamanca”».). La práctica de mantener un catálogo autoridades, unificando los puntos de acceso a información normalizada se la llama control de autoridades (Texidor, «Control de autoridades».). Este proceso tiene como objetivo:

- Probar que dos valores son idénticos comparando por la clave de las autoridades.
- Ayudar a completar metadatos con valores correctos.
- Mejorar la calidad de los metadatos.
- Mejorar la interoperabilidad compartiendo un nombre de autoridades con otra aplicación.
- Reducir los tiempos de carga de metadatos
- Organizar la información.

Un registro de autoridad puede contar con manejo de variantes, es decir variaciones sintácticas de un registro dadas por deletreos y faltas de ortografía, mayúsculas frente a variantes en minúsculas, fechas diferentes, etc. Por ejemplo en el catálogo de autoridades de la Biblioteca Nacional Mariano Moreno, la autoridad que representa a Rene Favaloro posee, entre otros, los campos ‘Nombre personal’, ‘Lugar asociado’, ‘Campo de actividad’, ‘Ocupación’, ‘Sexo’, etc, además de las siguiente variantes:

- Favaloro, René, 1923-2000
- Favaloro, René Gerónimo, 1923-2000
- Favaloro, René G. (René Gerónimo), 1923-2000

y su correspondiente ID en el sistema, el 000036235. La URI del registro mencionado es la siguiente
http://catalogo.bn.gov.ar/F/?func=direct&doc_number=000036235&local_base=BNA10

Ejemplos

VIAF

VIAF es un proyecto conjunto de varias bibliotecas nacionales, implementado y alojado por OCLC. El objetivo del proyecto es disminuir el coste e incrementar la utilidad de los ficheros de autoridad de las bibliotecas mediante la comparación y la correspondencia entre los ficheros de autoridades de las bibliotecas nacionales, y poner esa información disponible en Internet (OCLC, «VIAF»).

Por ejemplo, en el catálogo de autoridades de VIAF, la entrada que hace referencia a René Favaloro, se identifica con la URI https://viaf.org/viaf/94258659/ y las siguiente variantes.

- Favaloro, René G., 1923-2000
- Favaloro, René Geronimo
- Favaloro, René G., 1923-
- René Favaloro Médico argentino
- Favaloro, René, 1923-

CANTIC

El catálogo de nombres y títulos de autoridades de Cataluña (CANTIC) es un catálogo de autoridades cooperativas realizado en el Catálogo de Universidades de Cataluña (CCUC) y dirigido por la Biblioteca de Cataluña. (CANTIC, «CANTIC: Què és»).

Su objetivo es normalizar los puntos de acceso de los catálogos bibliográficos, mejorar la comunicación entre los distintos catálogos y, sobre todo, fomentar la búsqueda y recuperación de la información.

Sus registros están disponibles para las autoridades siguientes tipos:

- Nombres de individuos y familias.
- Nombres de congresos.
- Títulos uniformes.

Autoridades Museo São Paulo

La base de autoridades del Museo de São Paulo (MASP[1]) contiene nombres estandarizados de artistas plásticos, historiadores, críticos nacionales y extranjeros, junto con sus variaciones de nombres, fecha y lugar de nacimiento y muerte, área de actuación entre otras informaciones (Biblioteca do Museu de Arte de São Paulo (MASP), «Controle de Autoridades | BARTOC.org»). Contiene también, los nombres estandarizados entidades

[1] "Controle De Autoridades." Controle De Autoridades | BARTOC.org. http://bartoc.org/en/node/18610.

colectivas como museos, galerías, fundaciones entre otras instituciones relacionadas con las artes visuales.

DBLP

DBLP es un sitio web que posee referencias bibliográficas de artículos relacionados con las ciencias de la computación incluyendo información sobre autores, conferencias, series y revistas. Este catálogo indexa los metadatos bibliográficos básicos de las publicaciones académicas, y proporciona un hipervínculo a las ediciones electrónicas oficiales en las bibliotecas digitales de los editores. No hay un sistema de administración de bases de datos detrás de dblp. La información se almacena en millones de archivos. (DBLP, «dblp: What is dblp?»)

DBLP dispone de una página de búsqueda destinada a ayudar a los usuarios a encontrar perfiles de autor, conferencias, publicaciones periódicas o publicaciones individuales en la base de datos. Otra forma de realizar búsquedas es a través de una API (DBLP, «dblp: How to use the dblp search API?») que cuenta con tres servicios proporcionados para buscar publicaciones, personas (autores/editores) y uno para lugares de desarrollo (revistas/conferencias/etc). Donde:

- http://dblp.org/search/publ/api para consultas de publicación
- http://dblp.org/search/author/api para consultas de autor
- http://dblp.org/search/venue/api para consultas sobre lugares de desarrollo

Este catálogo puede ser descargado en formato XML junto con el archivo DTD que valida la estructura del mismo.

BARTOC

El Registro de Tesauros, Ontologías y Clasificaciones de Basilea (BARTOC) (BARTOC, «About | BARTOC.org».) es un catálogo de autoridades desarrollado en la Biblioteca de la Universidad de Basilea, publicado en noviembre de 2013. Su objetivo principal es documentar sistemas de organización del conocimiento (KOS), como clasificaciones, tesauros y archivos de autoridad, en un solo lugar para lograr una mayor visibilidad, resaltar sus características, hacer que sean buscables y comparables, y fomentar el intercambio de conocimientos.

El sitio web y la base de datos BARTOC se basan en Drupal. Su contenido está disponible como dominio público con la Dedicación y licencia de dominio público (PDDL).

Además de poder consultar el catálogo de autoridades a través del sitio web, los registros de BARTOC pueden ser recuperados en formato RDFa como Linked Open Data a partir de

su URI. La URI de cada registro representa la clave de la autoridad que representan, teniendo la URI un formato similar a http://bartoc.org/en/node/{ID} donde {ID} es un número.

Otro formato disponible de las autoridades en BARTOC es JSKOS, el cual define una estructura de JSON para estructurar sistemas de organización del conocimiento (KOS). JSKOS admite la representación de conceptos, esquemas conceptuales, ocurrencias conceptuales y mapeos conceptuales con sus propiedades comunes.

También es posible descargar el catálogo en formatos XML, CSV, DOC, TXT, XLS y JSON.

C. Taxonomías

Una taxonomía es un sistema de clasificación que permite agrupar un conjunto de elementos dentro de categorías predefinidas, llamadas taxones («About Taxonomy».). Los taxones deben poder ser identificados unívocamente dentro de la taxonomía, pueden estar contenidos entre sí, o relacionadas de cualquier otra manera. Un ejemplo de taxonomía es la taxonomía biológica, utilizada para clasificar los seres vivos en reinos, especies y razas.

En los sistemas de gestión de contenido (CMS), como Drupal y Wordpress, las taxonomías tienen un rol fundamental ya que permiten organizar el contenido en categorías y subcategorías, en el caso de Wordpress. En Drupal las taxonomías se utilizan no solo para organizar contenido, sino como base arquitectural para el desarrollo de otros módulos, como pueden ser menús («Organizing Content with Taxonomy».).

Ejemplos

Universal Decimal Classification (UDC)

La Clasificación Decimal Universal (UDC) es una vocabulario controlado que tiene como finalidad representar todas las ramas del conocimiento humano organizandolas en una taxonomía. El UDC está estructurado de tal manera que los nuevos desarrollos y nuevos campos de conocimiento pueden incorporarse fácilmente debido a su disposición jerárquica permite. El código en sí es independiente de cualquier lenguaje o secuencia de comandos en particular y las descripciones de las clases correspondientes han aparecido en muchas versiones traducidas («UDC Consortium - About UDC».).

La característica más influyente de UDC es su capacidad para expresar no solo temas simples sino también relaciones entre sujetos. Esta función se agrega a una estructura jerárquica, en la que el conocimiento se divide en diez clases, luego cada clase se subdivide en sus partes lógicas, cada subdivisión se subdivide, y así sucesivamente. Cuanto más detallada sea la subdivisión, más largo será el número que la representa. La siguiente tabla

muestra cómo se divide el área del conocimiento dedicado a matemáticas y ciencias naturales con sus correspondientes notaciones.

Notación	Descripción
5	MATEMÁTICAS. CIENCIAS NATURALES
53	Física
539	Naturaleza física de la materia
539.1	Física nuclear. Física atómica. Física molecular

Tabla 1. Ejemplo de subdivisión en UDC

Además de poder navegar esta taxonomía a través de un navegador este vocabulario se encuentra publicado en formato SKOS para ser integrado a Linked Data así como es tambíen es posible descargarlo en formato XML/RDF («UDC Summary Linked Data»).

D. Tesauros

Un tesauro es una lista de términos controlados utilizados para representar conceptos. Por su estructura, es un vocabulario controlado y dinámico de términos relacionados semántica y jerárquicamente, los cuales cubren un dominio específico del conocimiento. (Carrrascosa, «Tesauros y ontologías»). Los términos que conforman el tesauro se relacionan entre sí pudiendo ser estas relaciones jerárquicas, de afinidad o preferenciales.

Ejemplos

AGROVOC

AGROVOC es un tesauro multilingüe que abarca todos los ámbitos de interés de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), entre ellos la alimentación, la nutrición, la agricultura, la pesca, las ciencias forestales y el medio ambiente («About AGROVOC | Agricultural Information Management Standards (AIMS)»).
AGROVOC permite la edición de tesauros multilingües y recursos RDF-SKOS en colaboración y puede ser accedido a través de un endpoint SPARQL así como puede ser descargado en formato RDF y NT o puede ser consultado en línea.
Cada versión de AGROVOC está disponible de dos formas:

1. AGROVOC Core: Únicamente el tesauro, con el registro de la fecha y hora de la liberación de la versión

2. AGROVOC LOD: Disponible en formato RDF/SKOS-XL, a través de un endpoint SPARQL. («AGROVOC Linked Open Data | Agricultural Information Management Standards (AIMS)».)

Getty Arts and Architecture Thesaurus

Getty Research Institute dispone de 4 vocabularios controlados con terminología y otra información sobre objetos, artistas, conceptos y lugares importantes para las diversas disciplinas que se especializan en arte, arquitectura y cultura («Art & Architecture Thesaurus (Getty Research Institute)»)

Los vocabularios controlados del Getty Resarch institute son:

- The Art & Architecture Thesaurus ® (AAT): incluye términos, descripciones y otros metadatos para conceptos genéricos relacionados con el arte, la arquitectura, la conservación, la arqueología, y otros patrimonios culturales. Se incluyen los tipos de trabajo, estilos, materiales, técnicas y otros.
- Getty Thesaurus of Geographic Names ® (TGN): contiene nombres, descripciones y otros metadatos para ciudades, imperios, sitios arqueológicos, y las características físicas importantes para la investigación del arte y la arquitectura. TGN podría estar relacionada con GISs, mapas y otros recursos geográficos.
- Union List of Artist Names ® (ULAN): nombres, biografías, personas relacionadas y otros metadatos sobre artistas, arquitectos, empresas, estudios, museos, clientes, y otras personas y grupos involucrados en la creación y el estudio del arte y la arquitectura.
- Cultural Objects Name Authority ® (CONA): está conformada por títulos, atribuciones, temas representados, y otros metadatos acerca de las obras de arte, arquitectura, y patrimonio cultural, tanto existentes como histórico.

Los datos de los vocabulario que dispone Getty se puede acceder desde:

- El sitio web de Getty: Los usuarios pueden buscar términos y nombres individuales en los vocabularios Getty a través de una interfaz web.
- Formatos Linked Open Data: Disponibles en JSON, RDF, N3/Turtle, y N-Triples a través de un endpoint SPARQL(«Getty Vocabularies»).
- Web services: Los vocabularios puede ser accedidos a través de los protocolos disponibles son SOAP 1.1 and 1.2, HTTP GET y HTTP POST. Las URLs donde se definen los webservices son las siguientes:
    - http://vocabsservices.getty.edu/AATService.asmx
    - http://vocabsservices.getty.edu/ULANService.asmx
    - http://vocabsservices.getty.edu/TGNService.asmx

- Tablas relacionales y XML: Es posible descargar los datos a modo de ejemplo, y recrearlos en una base de datos relacional. Esta metodología ésta próxima a ser discontinuada.

STW Thesaurus for Economics

STW es un tesauro que provee vocabulario sobre cualquier tema económico, conformado por casi 6.000 encabezados de materia estandarizados y aproximadamente 20.000 términos de entrada adicionales para respaldar palabras clave individuales («STW Thesaurus for Economics: Home»). También es posible encontrar términos técnicos usados en leyes, sociología o política, y nombres geográficos.

EuroVoc

EuroVoc es un tesauro multilingüe y multidisciplinario que abarca la terminología de los ámbitos de actividad de la Unión Europea. Contiene términos en 23 lenguas oficiales de la UE (alemán, búlgaro, checo, croata, danés, eslovaco, esloveno, español, estonio, finés, francés, griego, húngaro, inglés, italiano, letón, lituano, maltés, neerlandés, polaco, portugués, rumano y sueco), así como en tres lenguas de países candidatos a la adhesión a la UE: македонски (mk), shqip (sq), y српски (sr). («EuroVoc»)

Utilizan el tesauro EuroVoc, entre otros, el Parlamento Europeo, la Oficina de Publicaciones, parlamentos nacionales y regionales de toda Europa, administraciones nacionales y usuarios privados tanto de los países miembros de la UE como de terceros países.

E. Ontologías

Una ontología es una definición formal de tipos, propiedades, y relaciones entre entidades cuyo fin es representar el conocimiento (entidades, ideas, eventos, etc) limitando la complejidad del mismo a dominio en particular. Es considerado un vocabularios controlado ya que define los términos y las relaciones para la compresión de un área del conocimiento, así como las reglas para poder combinar los términos para definir las extensiones (Lapuente y Lapuente, «Ontologías».)

Las ontologías constan de términos (o clases), relaciones, propiedades, instancias y axiomas.

Los términos son un conjunto de objetos (físicos, tareas, funciones, etc.) y las interacciones entre estos son representadas por las relaciones. Las relaciones más utilizadas son:

- Hiponimia: palabra cuyo significado está incluido en el de otra. Relación es-un (is-a). Por ejemplo: la palabra ‘escritorio’ es hipónimo de ‘mueble’ (siendo mueble

hiperónimo de ‘escritorio’), ya que posee todos los rasgos semánticos y añaden otras características para diferenciarlas de esta.

- Meronimia: una palabra que nombra una parte de un todo, relación part-of(parte-de). Por ejemplo: la palabra ‘dedos’ es un merónimo de ‘manos’ siendo ‘manos’ holónimo de ‘dedos’.
- Sinonimia: relación que asocia dos términos que palabras que tienen un significado similar o idéntico entre sí. Por ejemplo, vehículo es un sinónimo de automóvil.

Los objetos que conforman una ontología se describen por medio de un conjunto de características o atributos llamados propiedades. Las instancias son la representación de los objetos. Por último los axiomas son teoremas que permiten definir aserciones que se cumplen siempre. Existen tres tipos de axiomas: relacionales, no-relacionales y generales (Carrrascosa, «Tesauros y ontologias».).

Ejemplos

Friend of a Friend (FOAF)

FOAF es una ontología compuesta por términos que describen personas, sus actividades y sus relaciones con otras personas y objetos («FOAF Vocabulary Specification»). Se centra principalmente en la existencia de personas en el mundo virtual, con muchas propiedades relacionadas con la actividad o identidad en línea como lo son: foaf:mbox, foaf:skypeID.

En la figura 1 se puede apreciar un ejemplo de instancia de foaf:Person, que representa a un recurso, identificado con la URI http://ejemplo/Juan, de tipo persona con nombre Juan y mail juan@ejemplo.com.

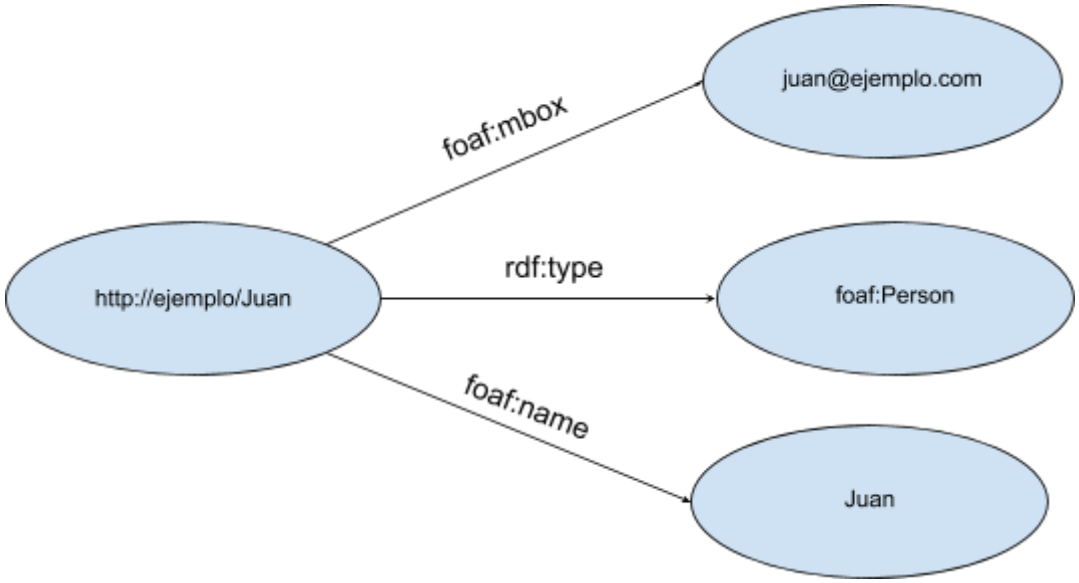

Figura 1. Representación de una instancia de foaf:Person

Socially Interconnected Online Communities (SIOC)

Esta ontología se usa para describir comunidades web como foros, blogs, listas de correo, wikis, etc. Es compuesta por 17 clases, 61 propiedades, y 25 propiedades de tipos de datos. Complementa a FOAF al hacer hincapié en la descripción de los productos de esas comunidades (publicaciones, respuestas, hilos, etc.). («SIOC Core Ontology Specification»)

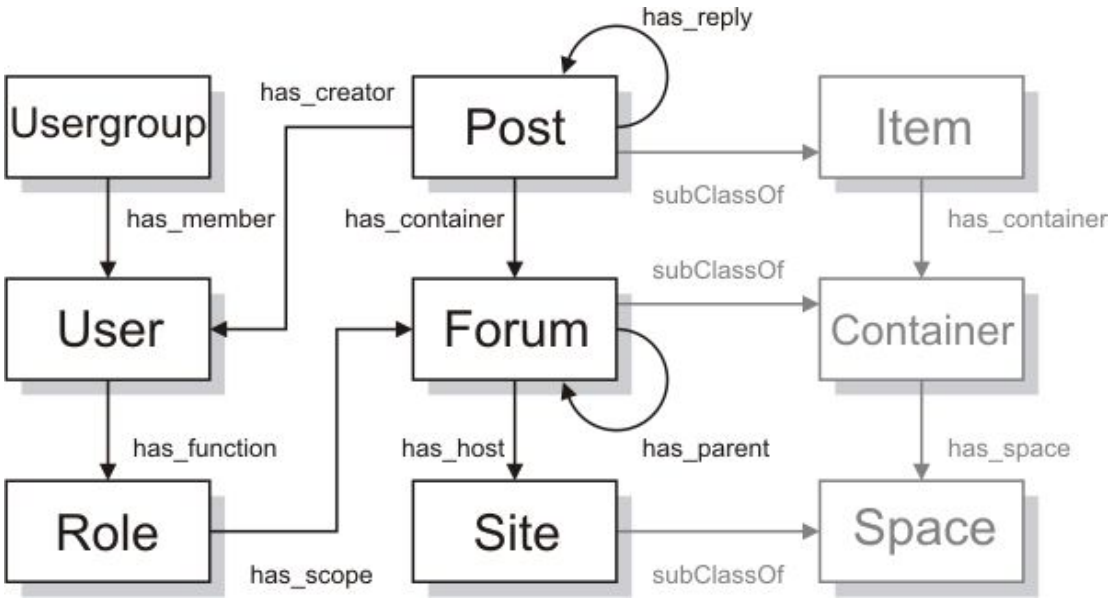

Figura 2. Ejemplo de entidades y relaciones definidas en SIOC («SIOC Core Ontology Specification»).

```
<sioc:Post rdf:about="http://blog.example.org/post/7#post">
 <dc:title>Post de ejemplo</dc:title>
 <dcterms:created>2008-05-25T09:33:30Z</dcterms:created>
 <sioc:has_creator>
  <sioc:User rdf:about="http://blog.example.org/user/pablo">
   <sioc:name>Pablo</sioc:name>
  </sioc:User>
 </sioc:has_creator>
 <sioc:content>...</sioc:content>
 <sioc:has_container rdf:resource="http://blog.example.org/#blog"/>
 <sioc:has_reply>
  <sioc:Post rdf:about="http://blog.example.org/post/7#comment1" />
 </sioc:has_reply>
 <sioc:topic rdf:resource="http://blog.example.org/blog/category/blogs/"/>
 <rdfs:seeAlso rdf:resource="http://blog.example.org/post/7.rdf" />
</sioc:Post>
```

Este ejemplo representa a un objeto sioc:Post identificado con `http://blog.example.org/post/7#post` que tiene las siguientes propiedades:

- dc:title con el valor "Posteo de ejemplo".
- dcterms:created con el valor de fecha del posteo.
- sioc:has_container con un objeto identificado como `http://blog.example.org/#blog` al que pertenece esta publicación
- sioc:has_creator representando el autor del posteo identificado con `http://blog.example.org/user/pablo`
- sioc:content representando el contenido del posteo con texto.
- sioc:topic que indican el tema del post como "Blogs" identificados como `http://blog.example.org/#blog`
- sioc:has_reply (que referencia a los comentarios de una publicación) identificado como `http://blog.example.org/post/7#comment1`

SCHEMA.ORG

Schema.org es una iniciativa de Google, Yahoo, Yandex y Microsoft que tiene como finalidad ayudar a los motores de búsqueda a interpretar el contenido de los sitios web. Para esto schema.org ha creado esquemas para representar entidades como organizaciones, personas, lugares, productos, películas, libros, eventos, reviews, etc.

Por ejemplo la clase Person, con identificador https://schema.org/Person, tienen propiedades como address para representar la dirección de residencia de una persona y puede tomar valores como texto libre estar asociado a la clase PostalAddress, definida para representar direcciones físicas con propiedades como addressCountry, utilizada para asociar la dirección con un país.

PREMIS OWL Ontology

La ontología PREMIS OWL fue diseñada para definir las entidades y las propiedades que se describen (Objetos, Eventos, Agentes y Derechos) en el Diccionario de datos PREMIS para la preservación de metadatos. Se encuentra implementado en RDF haciendo al mismo compatible con LinkedData y poder ser consultado a través de SPARQL. (Congress y Committee, «PREMIS OWL Ontology (PREMIS, Preservation Metadata Maintenance Activity, Library of Congress)»)

Actualmente se encuentra en la versión 3.0. A diferencia de la versión 2.0, se han incorporando mejores prácticas de Linked Data y conexiones a otras ontologías RDF como los términos de metadatos de Dublin Core y otros vocabularios de preservación. El grupo de

trabajo alienta a las personas en las comunidades de preservación, metadatos y datos vinculados a revisar y proporcionar comentarios antes de que se finalicen.

Formas de representación de vocabularios

RDF

RDF es un modelo de datos de la Web Semántica, diseñado para representar recursos en forma de tripletas sujeto-predicado-objeto, donde el sujeto indica el recurso representado, el predicado hace referencia a los rasgos del recurso, siendo este la relación entre el sujeto y el objeto. Este modelo de tripletas permite estructurar recursos formando un gráfico dirigido y etiquetado, donde las aristas representan el enlace entre dos recursos, presentados por los nodos del grafo. En la figura 3 se puede ver como se expresa que un recurso identificado con la uri http://ejemplo/perro que hace referencia a un perro es de un animal a través del recurso identificado con la URI http://ejemplo/animal.

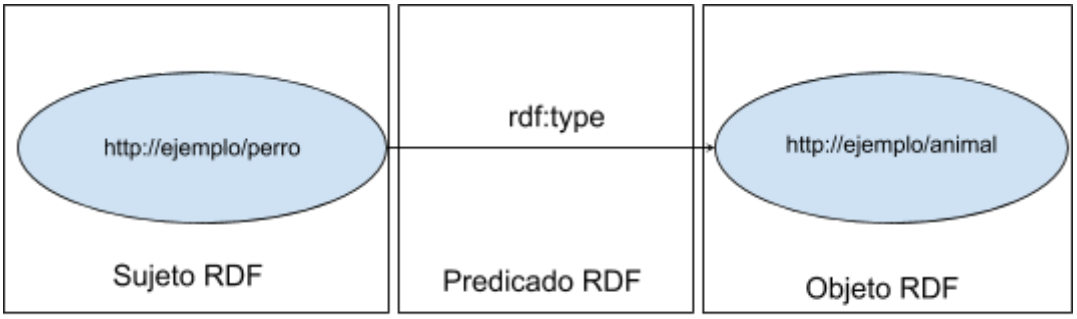

Figura 3. Ejemplo de un recurso representado en RDF

El uso de RDF también permite que los sistemas de organización del conocimiento se utilicen en aplicaciones de metadatos distribuidos y descentralizados. Los metadatos descentralizados se están convirtiendo en un escenario típico, donde los proveedores de servicios desean agregar valor a los metadatos recolectados de múltiples fuentes. («Learn RDF»). El uso de RDF permite obtener documentos en un formato que facilita su lectura por parte de aplicaciones informáticas, así como su intercambio y su publicación en la Web.
RDF puede ser serializado de varias formas, entre ellas podemos mencionar («RDF - Semantic Web Standards».)

- RDF/XML, primer estándar de serialización basada en XML.
- RDF/JSON, sintaxis basada en notación JSON.

- N3 o Notation3 es una forma abreviada de serialización no-XML de modelos en RDF con soporte para reglas basadas en RDF
- Turtle, subconjunto de N3 que solo puede serializar grafos RDF válidos.

Ejemplo de FOAF codificado en RDF:

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:foaf="http://xmlns.com/foaf/0.1/"
 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
 <foaf:Person>
  <foaf:name>Pedro</foaf:name>
  <foaf:mbox rdf:resource="pedro@ejemplo.com" />
  <foaf:knows>
   <foaf:Person>
    <foaf:name>Juan</foaf:name>
   </foaf:Person>
  </foaf:knows>
 </foaf:Person>
</rdf:RDF>
```

Este ejemplo representa a un objeto foaf:Person con las siguientes propiedades:

- foaf:name con nombre ‘Pedro’ que hace referencia al nombre de la persona.
- foaf:mailto con el valor `pedro@ejemplo.com`
- foaf:knows que representa la relación de conocimiento con otra persona con nombre ‘Juan’

OWL

OWL es un lenguaje diseñado para representar el conocimiento en la web semántica, a través de ontologías. OWL proporciona tres lenguajes diseñados para ser utilizado por diferentes usuarios donde cada uno de estos sublenguajes es una extensión de su predecesor más simple («Owl 101 - Cambridge Semantics | Cambridge Semantics».).
Los sublenguajes de OWL («Vista General del Lenguaje de Ontologías Web (OWL)») son:

- OWL Lite: diseñado para aquellos usuarios que necesitan principalmente una clasificación jerárquica y restricciones simples.
- **OWL DL:** diseñado para aquellos usuarios que quieren la máxima expresividad conservando completitud computacional (se garantiza que todas las conclusiones

sean computables), y resolubilidad (todos los cálculos se resolverán en un tiempo finito). OWL DL incluye todas las construcciones del lenguaje de OWL, pero sólo pueden ser usados bajo ciertas restricciones.

- **OWL Full:** dirigido a usuarios que quieren máxima expresividad y libertad sintáctica de RDF sin garantías computacionales.

SKOS

SKOS proporciona una forma estandarizada para representar en RDF la estructura básica y el contenido de esquemas conceptuales como listas encabezamientos de materia, taxonomías, esquemas de clasificación, tesauros y cualquier tipo de vocabulario controlado.
SKOS se ha diseñado para crear nuevos sistemas de organización de conocimiento o migrar los ya existentes adaptándolos a su uso en la Web Semántica. Proporciona un vocabulario muy sencillo y un modelo intuitivo que puede ser utilizado conjuntamente con OWL o de forma independiente («SKOS»).
En SKOS los conceptos se identifican con referencias URI. Estos conceptos pueden etiquetarse en cadenas de texto en uno o varios idiomas, documentarse y estructurarse a través de relaciones semánticas de diversa tipología. Este modelo permite mapear conceptos de diferentes esquemas, así como definir colecciones ordenadas y agrupaciones de conceptos. También permite establecer relaciones entre las etiquetas asociadas a los conceptos.

Capítulo 3 - Interoperabilidad en repositorios

Se conoce como interoperabilidad a la capacidad de interacción de un sistema con otros sin imponer dependencias evitando así restricciones de acceso o de implementación (Wolf, «Interoperabilidad: ¿A qué aspiramos cuando hablamos de ella? | SG Buzz».). Entre los motivos más importantes que se pueden mencionar sobre la importancia que tiene para un repositorio digital interactuar con otros sistemas se puede mencionar: la necesidad de integrarse con otros sistemas de la institución, ampliar el alcance y difusión de la producción intelectual local, incorporarse a redes regionales e internacionales para aumentar la visibilidad de su contenido, agilizar la ingesta de contenidos, normalizar la información expuesta mediante el cumplimiento de estándares, y más.

En el siguiente esquema se puede ver cómo el repositorio institucional interactúa con otros sistemas que no necesariamente son parte de la institución a la que pertenece

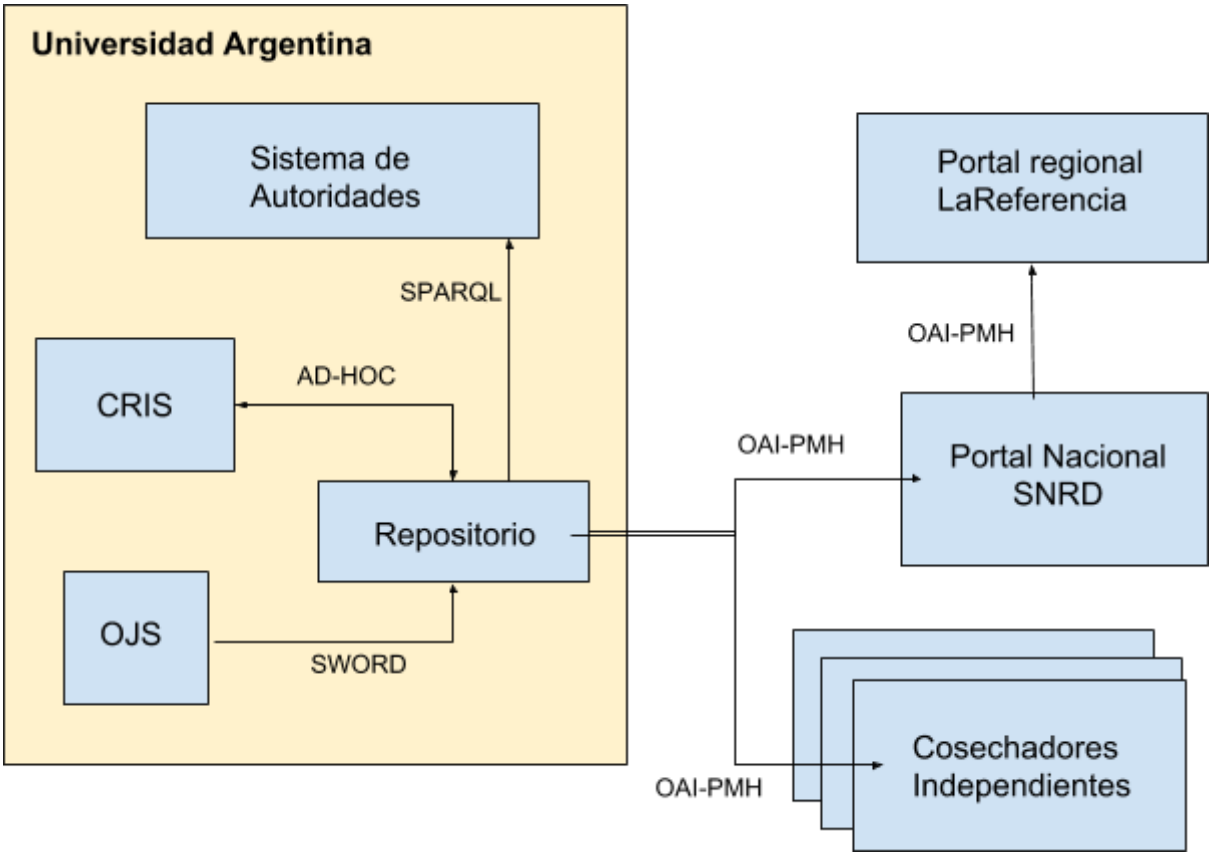

Figura 4. Interoperabilidad entre un repositorio y otros sistemas

Cada repositorio tiene necesidades distintas que conllevan a estructurar su información de forma tal que no necesariamente se corresponda con la estructura y los formatos utilizados en otros repositorios. Es por eso que para que exista interoperabilidad entre sistemas es necesario llevar la información hacia un formato común, minimizando su pérdida al integrar datos entre sistemas. A continuación se detallan algunos protocolos y tecnologías utilizados generalmente en el ámbito de los repositorios para interoperar.

OAI-PMH

El protocolo Open Archive Initiative-Protocol for Metadata Harvesting (OAI-PMH) es el utilizado para implementar interoperabilidad de metadatos en el ámbito de repositorios digitales. Este protocolo brinda un mecanismo simple para que lo repositorios interaccionen entre sí independientemente de la aplicación subyacente.
OAI-PMH define dos roles, uno de data provider, donde un repositorio puede recolectar o exponer metadatos, y otro denominado service provider, que aplica a los sistemas externos que recolectan los metadatos expuestos por los data providers, los procesan y los incorporan.

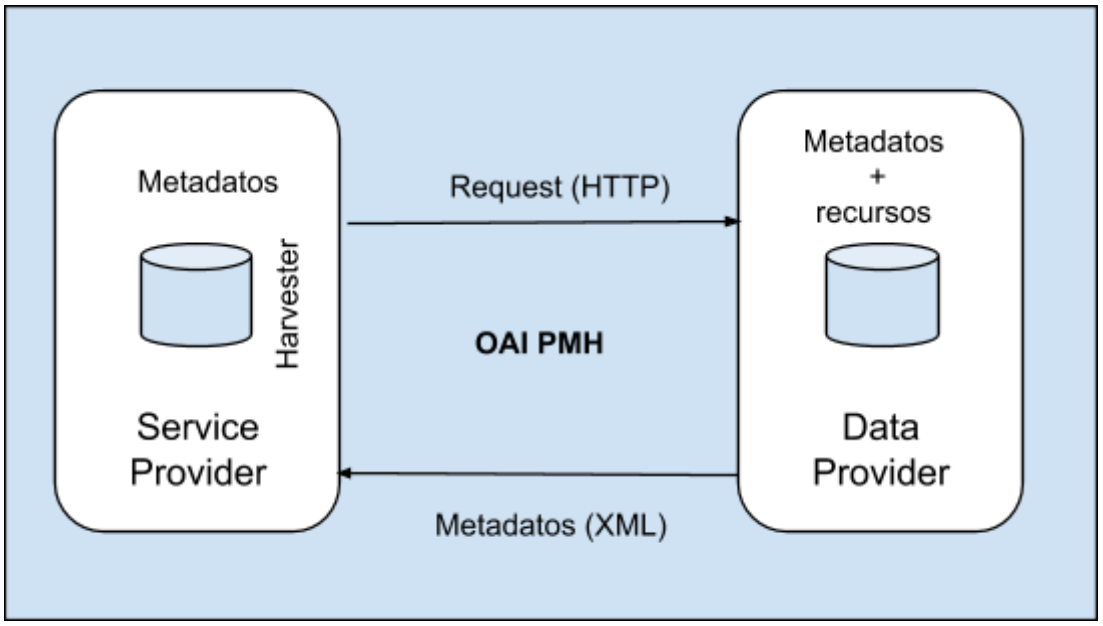

Figura 5. Protocolo OAI-PMH

Existen directrices de interoperabilidad sobre OAI-PMH que tienen como objetivo establecer un perfil de metadatos estandarizado que buscan asegurar la interoperabilidad de los repositorios. Por ejemplo en Argentina existe el Sistema Nacional de Repositorios Digitales (SNRD) cuyo objetivo es conformar una red interoperable de repositorios digitales a partir del establecimiento de políticas, estándares y protocolos comunes a todos los integrantes del sistema («Portal SNRD»).

SWORD

SWORD (Simple Web-service Offering Repository Deposit) es un estándar de interoperabilidad que permite enviar y recibir contenido desde múltiples fuentes. Esto permite enviar documentos desde otros sistemas, como por ejemplo un sistema de gestión de revistas científicas.

El protocolo SWORD funciona sobre HTTP, y está basado en el protocolo de publicación Atom. Un caso de uso de SWORD se da en el Portal de Revistas de la UNLP, donde un administrador de alguna revista de la UNLP envía al repositorio SEDICI un conjunto de artículos publicados en en su sitio web. Una vez recibido el artículo en SEDICI, un administrador verifica, completa los metadatos faltantes, y publica el nuevo número en el repositorio.

RSS

RSS se utiliza para la difusión de noticias y contenidos en línea. Se utiliza para el envío de información actualizada a usuarios que se han suscrito a la fuente de contenidos. El uso de RSS resulta de utilidad cuando se quiere que una web incluya, por ejemplo, un listado de los recursos cargados en el repositorio.

OpenSearch

OpenSearch es un conjunto de tecnologías que permiten compartir los resultados de una búsqueda permitiendo que otras aplicaciones y sitios web expongan contenidos del repositorio. Se integra fácilmente mediante RSS/Atom

Es una forma para que las páginas web y los motores de búsqueda publiquen sus resultados de forma accesible

Uso de Linked Data para interoperabilidad

La interoperabilidad semántica es un concepto que tiene sentido en el contexto de la Web Semántica, donde los recursos están representados de forma tal que puede ser procesados automáticamente con el objetivo de proporcionar servicios y funcionalidades de acceso avanzadas. (Narvaez y Piedra, «Un enfoque de Linked Data para garantizar la interoperabilidad semántica e integridad de datos académicos universitarios».)

Linked Data, o Datos Enlazados, es la forma que tiene la Web Semántica de vincular los datos distribuidos en la Web, extendiendola y permitiendo el procesamiento automático de

la información de un modo más exacto y completo (Berners-Lee, «Linked Data - Design Issues».).

La publicación de Datos Enlazados se fundamenta en cuatro principios básicos:

1. El uso de URIs para identificar los recursos de la Web.
2. Usar URIs-HTTP para que los usuarios puedan localizar y consultar estos recursos.
3. Ofrecer información útil acerca del recurso cuando la URI haya sido consultada, utilizando RDF para describir recursos y SPARQL para consultarlos.
4. Incluir enlaces a otras URIs relacionadas con los datos contenidos en el recurso, de forma que se potencie el descubrimiento de información en la Web.

Al nombrar los conceptos mediante URIs, se ofrece una abstracción del lenguaje natural y así se consigue evitar ambigüedades ofreciendo una forma estándar y unívoca para referirnos a cualquier recurso en la web.

Por ejemplo cuando debemos referirnos a un lugar por su topónimo éste puede variar en función del idioma, (Argentina, Аргентина, Arc'hantina, etc). Si usáramos el nombre para referirnos a los lugares, las múltiples acepciones que podría adoptar, dificultaría el tratamiento automatizado de la información.

De esta forma, si utilizamos un identificador único como http://dbpedia.org/resource/Argentina, cualquier aplicación podría hacer referencia al mismo lugar, independientemente de la ambigüedad del lenguaje natural.

En el núcleo de la Web Semántica se encuentran las ontologías, que son el medio para describir los recursos de la web. Las ontologías y los vocabularios abiertos forman el esquema base a partir del cual se describen los recursos y entidades de la Web. Los datos adquieren significado a través de ontologías, que se han ido construyendo gradualmente de acuerdo las necesidades de cada dominio de conocimiento.

Como dijimos anteriormente cada institución puede manejar diferentes formatos de datos o esquemas de metadatos o vocabularios y aunque OAI-PMH facilite el intercambio de metadatos entre repositorios aún pueden persistir problemas para integrar los datos extraídos. La reutilización de recursos ontológicos propuesto por el enfoque Linked Data es muy importante para incrementar el grado de interoperabilidad semántica entre sistemas, de manera que se construya un entorno en el que los agentes software pueden ejecutar colaborativamente tareas de procesamiento sobre los datos que actualmente se limitan a mostrar.

Capítulo 4 - Uso de vocabularios controlados y autoridades en DSpace

DSpace es un software de código abierto diseñado para la creación y gestión de repositorios digitales desarrollado y mantenido por una amplia comunidad de usuarios que permite almacenar y describir material digital, distribuir los recursos digitales de una organización a través de la web a través de un sistema de búsqueda y recuperación y preservar recursos digitales a largo plazo.

Este software posee un modelo de datos simple, con metadatos no jerárquicos e independencia de los formatos de archivos. Los repositorios basados en DSpace poseen una o más comunidades de nivel base que se organizan jerárquicamente en subcomunidades. Las comunidades pueden ser interpretadas como espacios de trabajo y las colecciones agrupan a los ítems. Los ítems del repositorio se componen de objetos digitales o bitstreams que pueden ser audios, PDFs, Documentos de texto, imágenes, videos, entre otros y que representan la obra en sí que se quiere publicar. Todo ítem está asociado a un conjunto de metadatos que describen a los recursos y sus bitstreams.

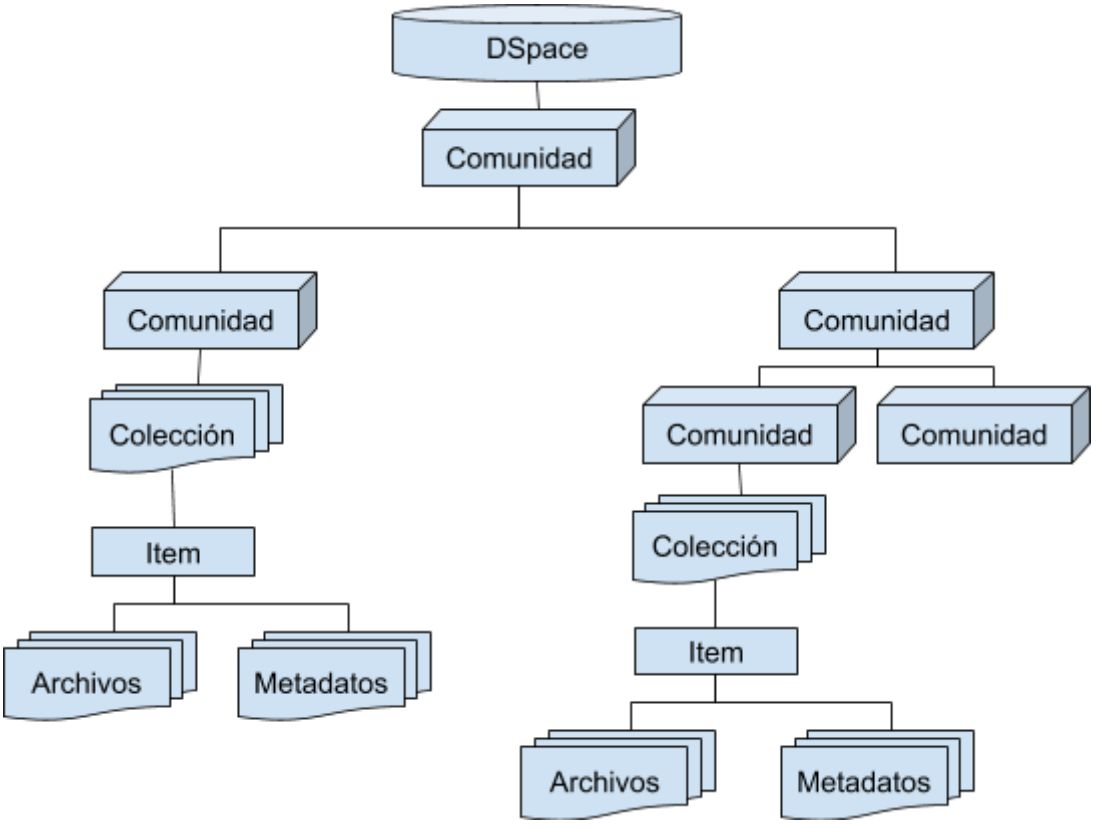

Figura 6. Organización y estructura de recursos digitales en DSpace

DSpace permite definir perfiles de metadatos a partir de la combinación de elementos de diferentes schemas, el uso de calificadores (schema.element.qualifier)

Arquitectura de DSpace

La arquitectura de DSpace se divide en 3 grandes grupos:

- Módulos de Aplicación: contiene componentes que se comunican con el mundo fuera de la instalación individual de DSpace, por ejemplo, la interfaz de usuario web y el protocolo OAI para el servicio de recolección de metadatos.
- Módulos de Lógica de negocio: se ocupa de la administración del contenido, los usuarios, la autorización y el flujo de trabajo
- Módulos de Almacenamiento: responsable del almacenamiento físico de metadatos y contenido

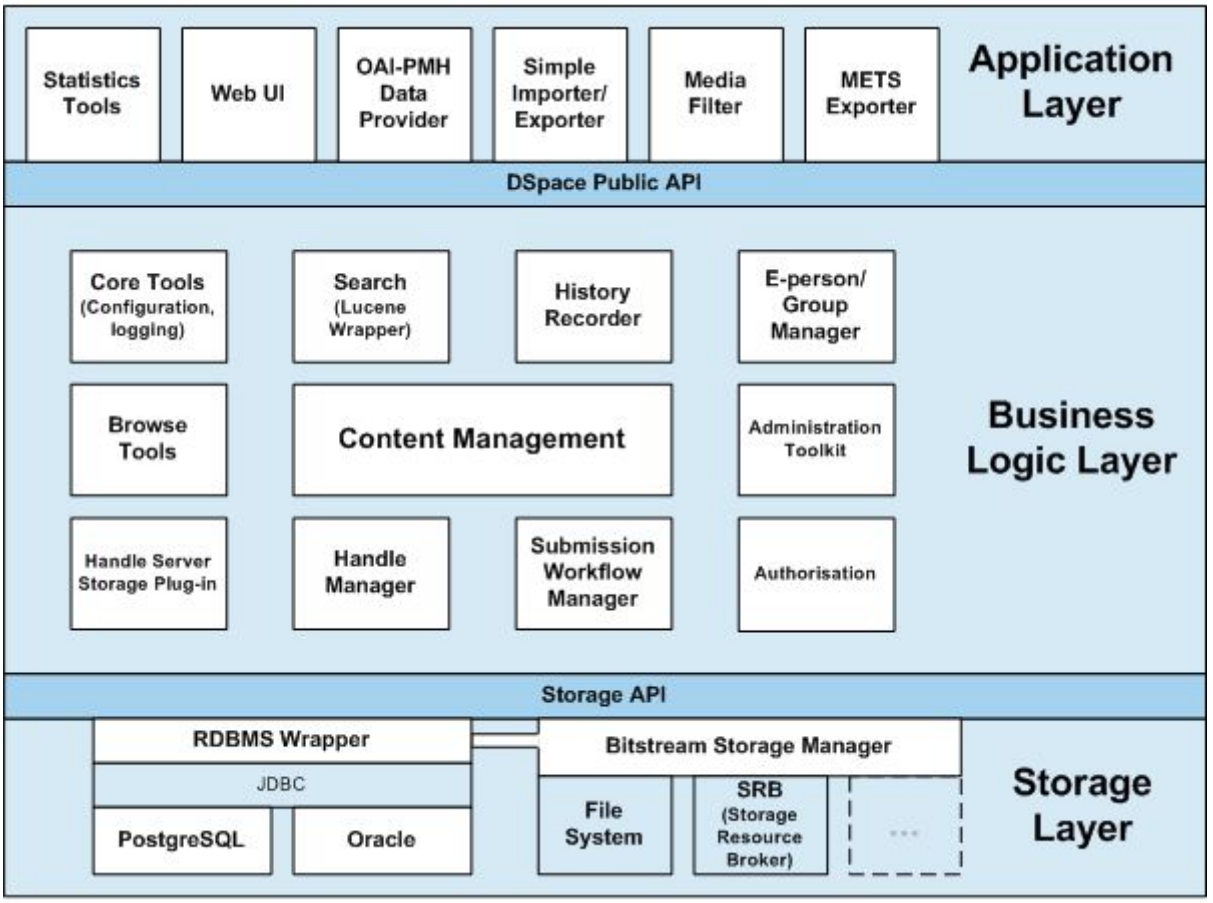

Figura 7 arquitectura de DSpace («Architecture - DSpace Documentation - DuraSpace Wiki».)

DSpace dispone de varios plugins (extensiones opcionales) para extender su funcionalidad, entre ellos:

- Plugins para la autenticación

- Plugins para el manejo de autoridades
- Plugins para procesamiento de los bitstreams (media filters)

Uso de autoridades en DSpace

Las primeras versiones de DSpace se concibieron inicialmente casi sin ningún tipo de control de calidad de metadatos. Luego, en versiones posteriores de DSpace, se incorporaron 3 formas de controlar los valores (DuraSpace, «Authority Control of Metadata Values - DSpace - DuraSpace Wiki».):

1. value-pairs: son asociaciones fijas de etiqueta-valor. Típicamente se usan para controlar vocabularios chicos y estáticos como tipología e idioma.
2. vocabularios controlados: permite conectar los metadatos con una jerarquía de valores almacenados en un archivo estático. Se usa típicamente para Sist. de clasificación predefinidos que no cambian con frecuencia.
3. módulo de autoridades: permite conectar los metadatos con entradas de autoridades externas. Es mucho más flexible, dinámico y desacoplado que las otras opciones, aunque requiere implementar un plugin.

A. Value-Pairs

Los elementos value-pairs permiten restringir los valores de metadatos. Este mecanismo no es considerado control de autoridades porque solo existe para este proceso y no afecta todas las fases de la gestión de metadatos.

Ejemplo: los value-pairs para el metadato *dc.language* a través de la propiedad *name* con el valor *common_iso_languages*.

```
<field>
   <dc-schema>dc</dc-schema>
   <dc-element>language</dc-element>
   <dc-qualifier>iso</dc-qualifier>
   <repeatable>false</repeatable>
   <label>Language</label>
        <input-type value-pairs-name="common_iso_languages">dropdown</input-type>
   <hint>Select the language of the main content of the item...</hint>
   <required></required>
</field>
```

Luego, en el mismo archivo XML se encuentran definidos los valores de common_iso_languages de la siguiente manera:

```
<value-pairs value-pairs-name="common_iso_languages" dc-term="language_iso">
```

B. Vocabularios controlados

Los vocabularios controlados en DSpace son definidos en archivos XML. Esta forma de definir los vocabularios es muy sencilla y permite definir una navegación jerárquica. Este enfoque es de utilidad en el caso de vocabularios controlados ya definidos y que no son propensos a cambios, ya que el hecho de tener que editar un archivo por cada vocabulario a controlar puede resultar engorroso. Por ejemplo, si queremos limitar los valores que puede tomar un metadato en particular a un conjunto de tipos de documento existente, lo definimos de la siguiente manera:

```
<?xml version="1.0" encoding="UTF-8"?>
<node id="sedici:types" label="">
<isComposedBy>
  <node id="sedici:types/articulo/articulo" label="Artículo"/>
  <node id="sedici:types/articulo/comunicación" label="Comunicacion"/>
  <node id="sedici:types/objetoDeAprendizaje/objetoDeAprendizaje" label="Objeto de
Aprendizaje"/>
</isComposedBy>
</node>
```

Por cada término controlado se define un label y un identificador que asociará con vocablo de forma transparente al usuario.

C. Módulo de autoridades

Utiliza una una arquitectura de plug-in que facilita la integración de nuevas autoridades sin modificar ningún código del núcleo de DSpace y poder controlar los metadatos que uno desee. Este módulo afecta todas las fases de la gestión de metadatos y puede aplicarse a cualquier metadato del esquema subyacente.

El módulo de autoridades fue diseñado para controlar dos aspectos básicos llamados Choice Management y Authority Control.

Choice Managment

La carga de metadatos controlados por autoridad se sustenta en un submódulo denominado Choice Management, que básicamente brinda un mecanismo para la selección de los

posibles valores de un metadato provenientes de una base de autoridades, a partir de un valor de consulta.

Un ejemplo de uso de este módulo en DSpace se puede apreciar en el proceso de carga de metadatos para un ítem determinado, más precisamente cuando un usuario completa parcialmente un campo controlado. A partir de valores propuestos, el sistema ofrecerá un conjunto de valores posibles para ese campo como se puede ver a continuación.

Figura 8. Ejemplo de Choice Managment en DSpace

Clases involucradas en DSpace

Para integrar nuevas autoridades deben utilizarse las siguientes clases e interfaces:

1. *org.dspace.content.authority.Choice*: Esta clase representa un término controlado por una autoridad. No tiene lógica; es símplemente un contenedor que contiene los atributos authority, label, confidence y value.
2. *org.dspace.content.authority.Choices*: Representa el resultado de una búsqueda de términos de autoridad y se compone de una serie de objetos Choice candidatos, un entero que representa el nivel de confianza de los valores del conjunto retornado y otros datos sobre las cantidades total y parcial de términos disponibles.
3. org.dspace.content.authority.ChoiceAuthority: Esta interfaz, se define para tomar valores de una autoridad que puede ser provista de cualquier fuente, ya sea una base de datos local o no, un servicio prestado por un tercero, documentos en cualquier formato, algo que provea términos controlados.

La interfaz org.dspace.content.authority.ChoiceAuthority cuenta con 3 métodos:

- *getMatches*: retorna un objeto *Choices* con todos los valores de una autoridad que coincidan con un valor provisto por el usuario. Por otra lado, algunas autoridades con un pequeño conjunto de valores pueden simplemente devolver todo el juego completo para cualquier valor de la muestra.
- *getBestMatch*: tiene como objetivo retornar un única autoridad que "mejor coincida" con el valor ingresado. Este método se usa típicamente en ingresos de metadatos en modo no interactivo (en modo batch) donde no hay un agente interactivo para elegir entre las opciones.
- *getLabel*: retorna una etiqueta (*label*) visible para el usuario de una autoridad dada, que se indica a partir de su clave. Este método puede ser llamado varias veces al cargar una página Web por lo que debe ser implementado de forma eficientemente.

Authority Control

Este mecanismo trabaja junto con el módulo Choice Managment identificando una autoridad que contiene un valor elegido. A los valores de metadatos que tienen una autoridad asociada, se le agrega en la tabla *metadatavalue* de la base de datos de DSpace 2 campos extra:

1. *confidence*: que indica el nivel de confianza del authority.
2. *authority: que* identifica la autoridad proveniente de la fuente externa.

text_value	authority	Confidence
Universidad Nacional de La Plata (UNLP)	http://ejemplo/auth/node/86555	600
Cañueto, Matías F.	http://ejemplo/auth/node/204702	600
Attribution 4.0 International (BY 4.0)	http://creativecommons.org/licenses/by/4.0/	600
UNLP	http://ejemplo/auth/node/86555	500
Cañueto, Matías G.		-1

Cuadro x. Ejemplo de metadatos controlados por autoridad en base de datos.

El módulo también permite restringir la elección de los valores de metadatos vinculados con el módulo de autoridades en 2 modos:

- Abierto: para permitir la carga de valores que no provengan de la fuente de autoridades.
- Cerrado: obliga a que todo valor ingresado sea reconocido por la fuente de autoridades.

Capítulo 5 - Herramientas de gestión de Vocabularios Controlados y Autoridades

En este capítulo se hace una descripción de tres alternativas para desarrollar un sistema de gestión de vocabularios controlados para el repositorio institucional de la UNLP, SEDICI.
Este software debe ser personalizable, capaz de adaptarse a las necesidades propias del repositorio y debe poder importar los datos almacenados en el sistema de gestión de vocabularios anterior. Otra de las características que debe cumplir este sistema es la de poder interoperar de forma sencilla con el repositorio sin realizar cambios sustanciales en el mismo.
La elección de estas tres alternativas surge de observar las herramientas utilizadas para implementar los sistemas de gestión de vocabularios controlados detallados en el capítulo anterior, la disposición del código de la aplicación y su flexibilidad.

VocBench

VocBench es una plataforma pensada para trabajar de manera colaborativa en el desarrollo y mantenimiento de ontologías OWL, tesauros SKOS (XL) y conjuntos de datos RDF genéricos. Este sistema organiza el trabajo en proyectos, donde diferentes usuarios pueden colaborar acorde a los privilegios de edición que posean para cada uno. («VocBench: A Collaborative Management System for SKOS-XL Thesauri».)
El backend de este sistema se encuentra desarrollado sobre Semantic Turkey, una plataforma de servicios RDF para la Gestión y Adquisición del Conocimiento (Knowledge Management and Acquisition) realizada por ART Research Group de la Universidad de Roma "Tor Vergata".
La plataforma VocBench permite definir grupos de usuarios, como administradores del sistema, administradores de proyectos y hasta roles más especializados que tratan con aspectos más generales del modelado de datos.
Los administradores del sistema tienen la capacidad de configurar el sistema, crear proyectos, crear nuevos usuarios, nuevos roles y asignar usuarios a proyectos con un rol dado.
Los roles definen conjuntos de capacidades que se pueden asignar a usuarios existentes, por proyecto es decir que a un usuario dado se le puede asignar el rol de "Lexicógrafo" en un determinado proyecto, mientras que en otro puede tener un rol totalmente distinto, como puede ser "Gerente de proyecto”.

VocBench también ofrece un mecanismo de verificación para tesauros SKOS y SKOS-XL debido que el estándar SKOS incluye muchas limitaciones que no se expresan a través de axiomas OWL y que necesitan maquinaria dedicada para poder verificarse. Actualmente VocBench no aplica estos controles sobre ontologías OWL pero si planean hacerlo en un futuro. La verificaciones implementadas chequean cuestiones como la integridad de la estructura de los datos (jerarquía, relaciones, etc.), la validez de las etiquetas definidas para cada término y que las URIs de los vocabularios sean válidas.

Integración de vocabularios

Esta herramienta permite crear proyectos en SKOS y OWL para manipular tesauros y ontologías respectivamente así como también permite cargar vocabularios existentes.
Por ejemplo es posible cargar la ontología de FOAF creando un nuevo proyecto con URI base "http://xmlns.com/foaf/0.1/" e importando la ontología descargada en RDF desde http://xmlns.com/foaf/spec/index.rdf

La siguiente imagen muestra una captura de pantalla de la vista de la ontología FOAF importada

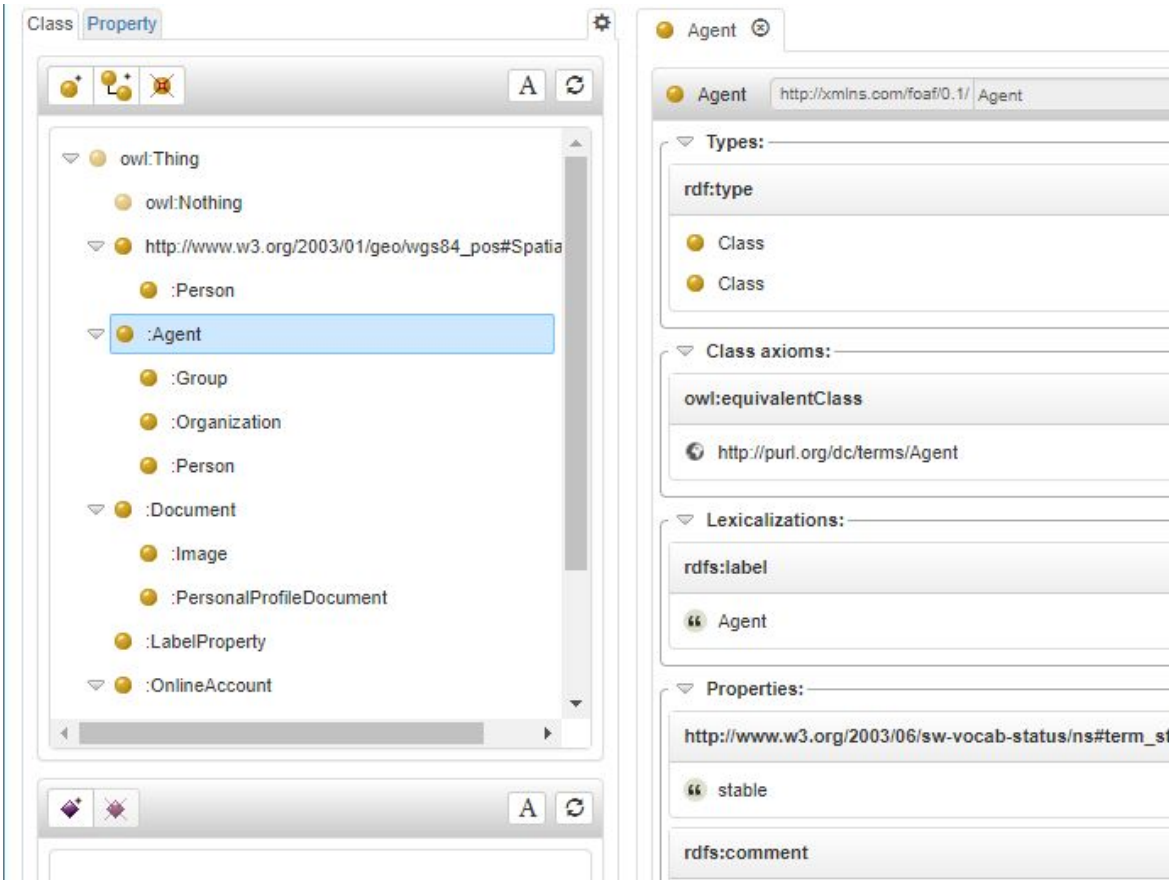

Figura 9. Vista de ontología importada en VocBench

Una vez importada la ontología, es posible crear un instancia del tipo *foaf:Person* como se muestra a continuación:

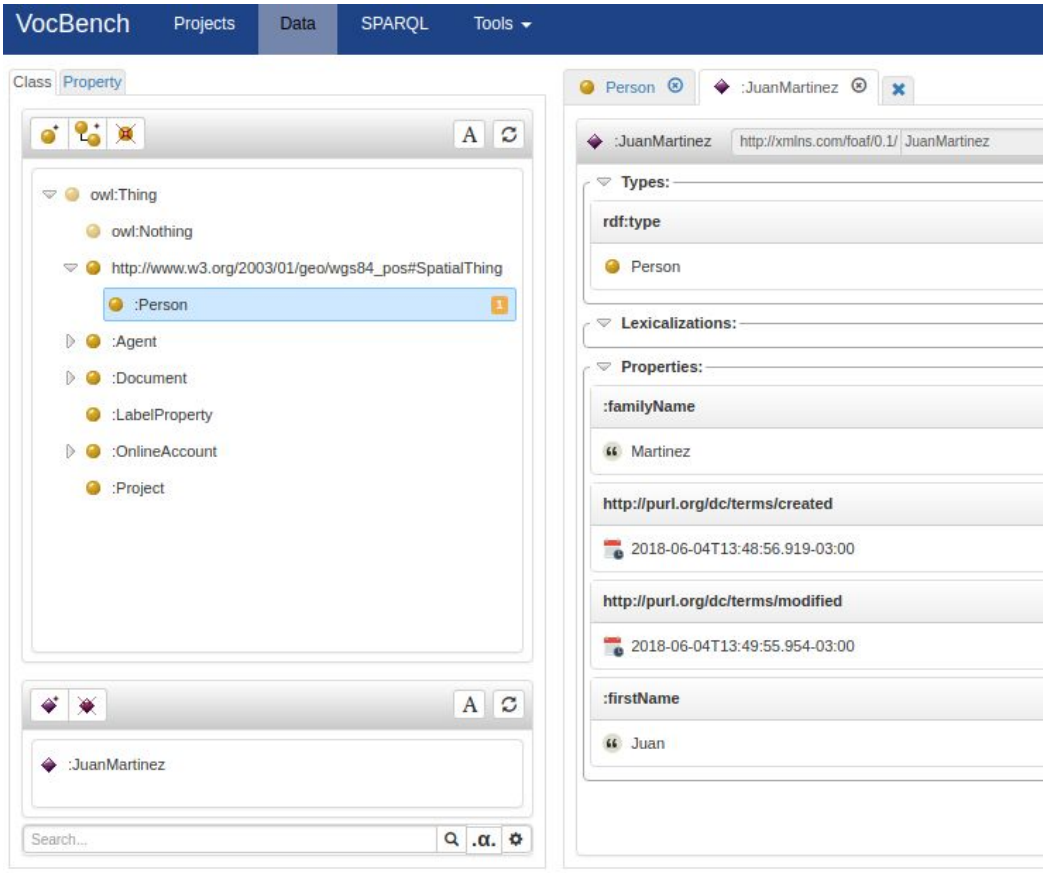

Figura 10. Creación de instancia del tipo foaf:Person

También es posible crear un proyecto SKOS para gestionar un tesauro grande conectándose a un almacén RDF externo, haciendo uso del historial, la validación y la inferencia de dicho tesauro, como puede ser importar el tesauro "Eurovoc". A continuación se muestra el tesauro Eurovoc importado en VocBench.

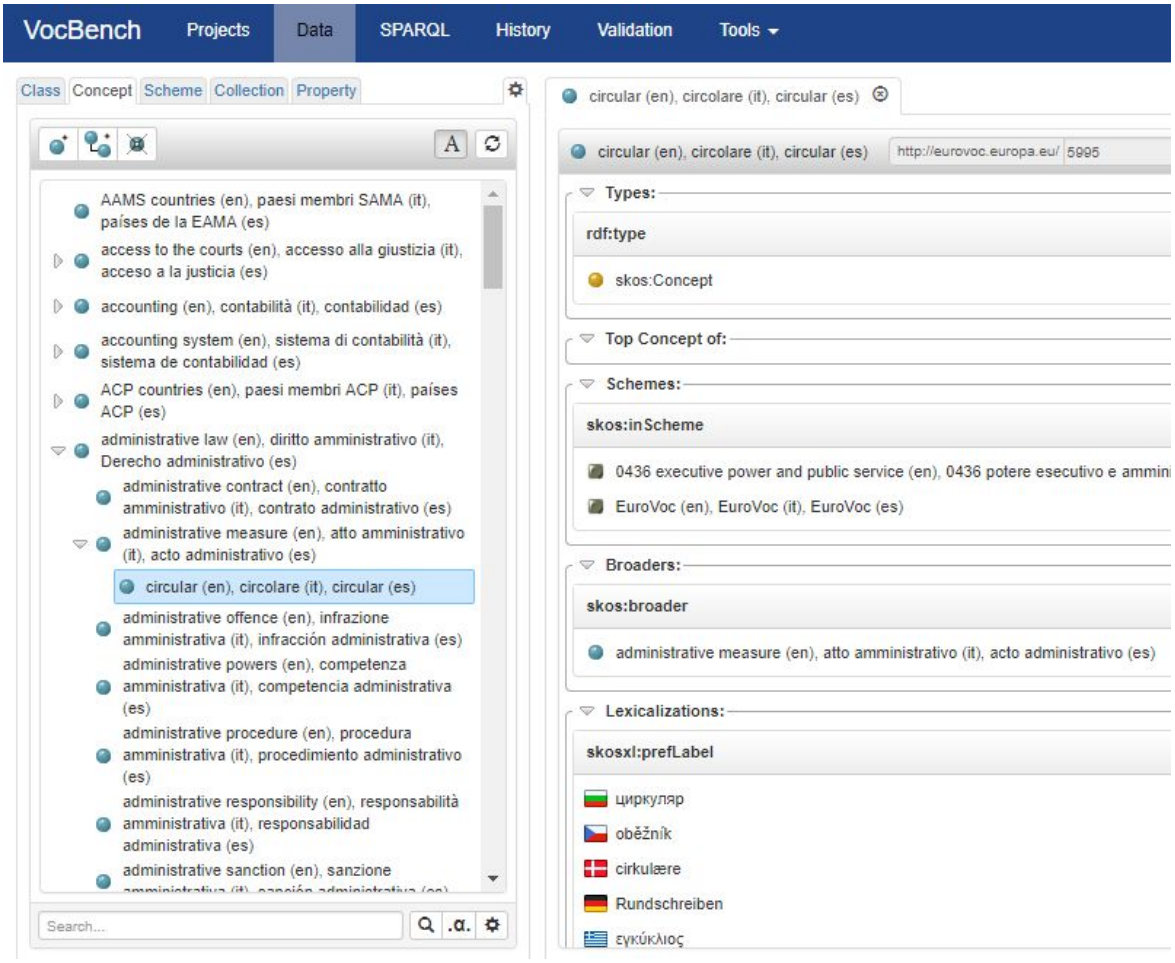

Figura 11. Vista del tesauro EuroVoc en VocBench

Interoperabilidad

VocBench dispone de un endpoint SPARQL desarrollado sobre YASGUI que permite publicar los vocabularios almacenados en formato, JSON, CSV, TSV, XLSX, ODS y RDF («YASGUI»). La siguiente captura de pantalla muestra como es la interfaz del endpoint SPARQL al que se le realiza una consulta pidiendo las propiedades firstName y familyName sobre los términos de tipo foaf:Person.

```
SELECT ?s ?name ?surname WHERE {
    ?s a <http://xmlns.com/foaf/0.1/Person> .
    ?s :firstName ?name .
    ?s :familyName ?surname
}
```

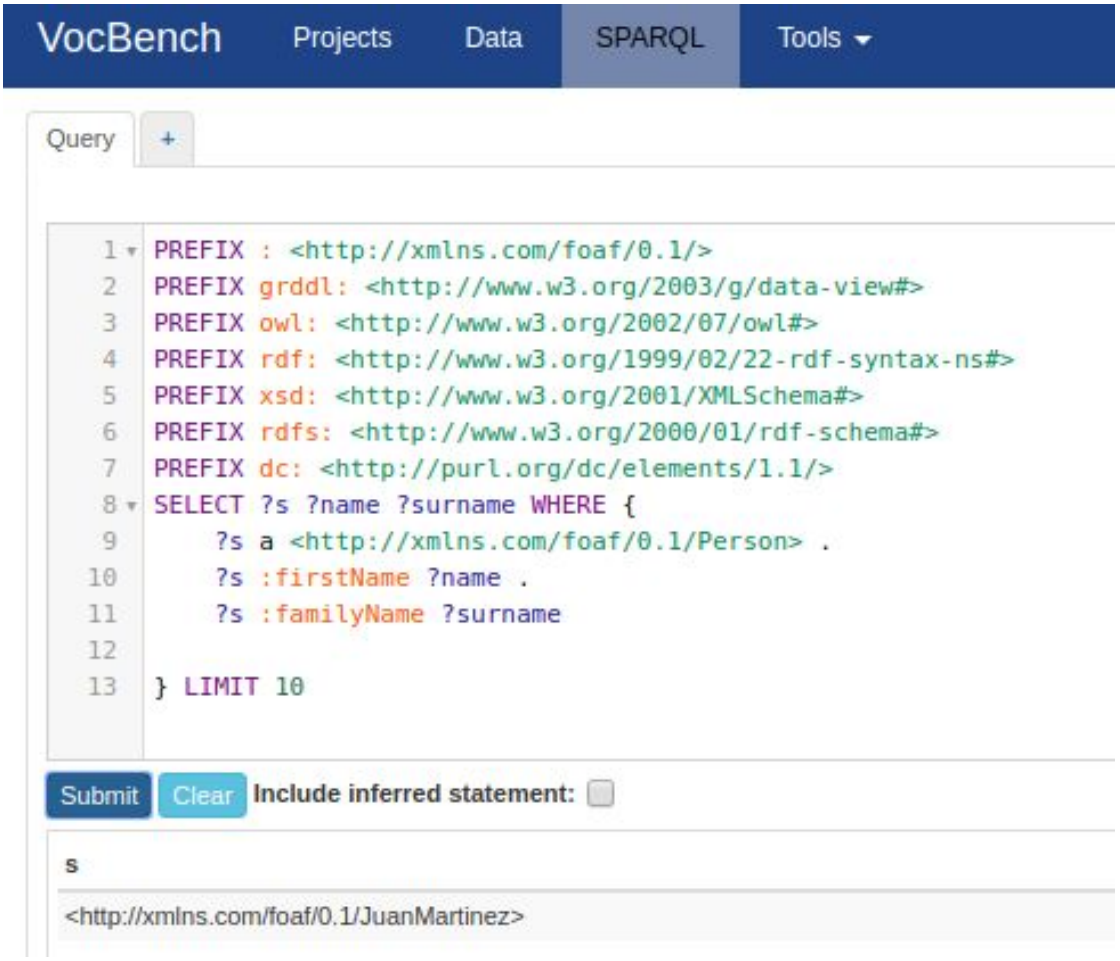

Figura 12. Consulta en el endpoint SPARQL en VocBench

La respuesta en formato JSON es la siguiente:


```
{
   "head":{ "vars":[ "s", "name", "surname" ] },
   "results":{
      "bindings":[
         {
            "s":{ "type":"uri", "value":"http://xmlns.com/foaf/0.1/JuanMartinez"},
            "surname":{ "type":"literal", "value":"Martinez" },
            "name":{"type":"literal", "value":"Juan"}
         }
      ]
   }
}
```

Casos de uso

La comunidad de usuarios de VocBench está creciendo y al día de hoy incluye a la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) que actualmente gestiona el tesauro AGROVOC, el Glosario de Biotecnología y otros metadatos bibliográficos. Otras organizaciones que utilizan VocBench son el proyecto data.fao.org, la Oficina de Publicaciones de la Comisión Europea, la Agencia Europea de Medio Ambiente y el Senado italiano

TemaTres

TemaTres es una herramienta Web que permite la creación y gestión de vocabularios controlados, tesauros, taxonomías y otros modelos de representación formal del conocimiento. TemaTres permite representar, publicar y exponer vocabularios controlados en varios esquemas de metadatos como pueden ser SKOS-Core, Dublin Core, MADS, entre otros. («TemaTres: servidor de vocabularios controlados».)
Esta herramienta dispone de un workflow de gestión de términos que establece qué puede hacerse con un término según sea su estado, y cuáles son las condiciones que debe reunir para poder pasar al siguiente. Los estados por los que pueden atravesar dichos términos son Candidato, Aceptado y Rechazado.

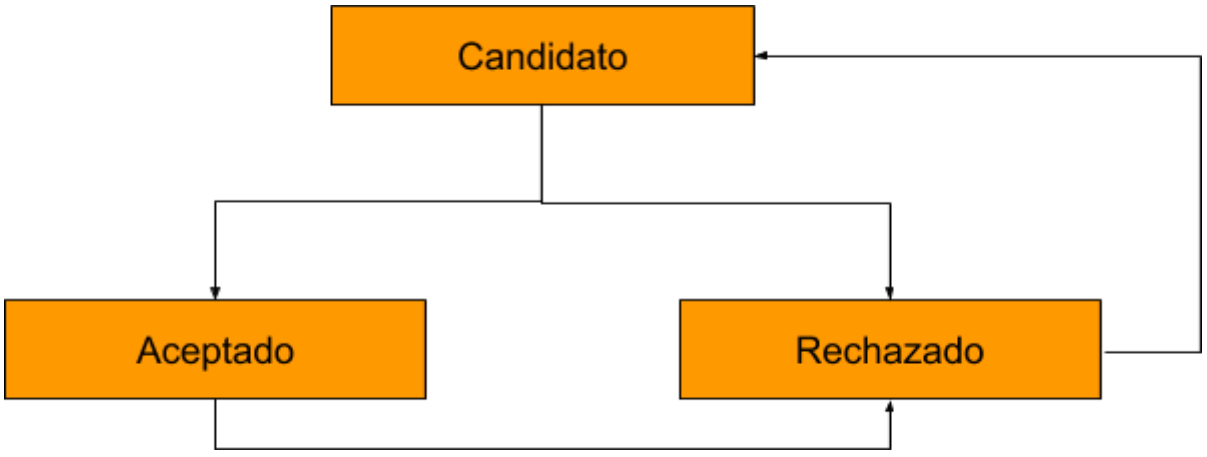

Figura 13. Posibles estados de un término en TemaTres

Un término puede ser eliminado en cualquiera de estos tres estados, salvo que el mismo se encuentre en aceptado y no sea término libre. Por otra parte, sólo los términos aceptados pueden tener relaciones con otros términos.
A continuación se adjunta una captura de pantalla donde se ve como es el alta de un término en TemaTres

Figura 14. Alta de término en TemaTres

También es posible representar las relaciones jerárquicas entre los términos, indicando si estos son Términos Genéricos (TG), Términos Específicos (TE), Términos No-preferido (UP) y Término Relacionado (TR) como se muestra en la siguiente imagen:

Figura 15. Alta de relación de términos en TemaTres

Figura 16. Vista de en detalle de un término en TemaTres

Integración de vocabularios

TemaTres dispone de dos modelos posibles para la implementación de vocabulario multilingüe:

- Modelo centrado con un vocabulario de referencia local
- Modelo federado con vocabulario de referencia externo

Modelo centrado con un vocabulario de referencia local

En este modelo TemaTres contiene y gerencia en el mismo vocabulario lo términos propios del vocabulario fuente y las relaciones de equivalencia hacia los términos de los vocabularios de destino.

Esta opción puede resultar útil si el vocabulario de destino no existe o sus servicios terminológicos no están disponible o las correspondencias terminológicas resultan incompletas.

En la siguiente imagen se muestra cómo relacionar dos términos locales

Vista de alta de equivalencia terminológica local

Modelo federado con vocabulario de referencia externo

A través de este modelo, TemaTres establece relaciones entre distintos vocabularios independientes. Para esto los vocabularios destino deben ser capaces de ofrecer web services terminológicos según el patrón de servicios de TemaTres para que el vocabulario de origen busque términos en el vocabulario de destino y establezca las relaciones entre los términos de cada uno de los vocabularios.

En la siguiente imagen se muestra cómo se puede relacionar el término África con algún términos del vocabulario del tesauro de la UNESCO

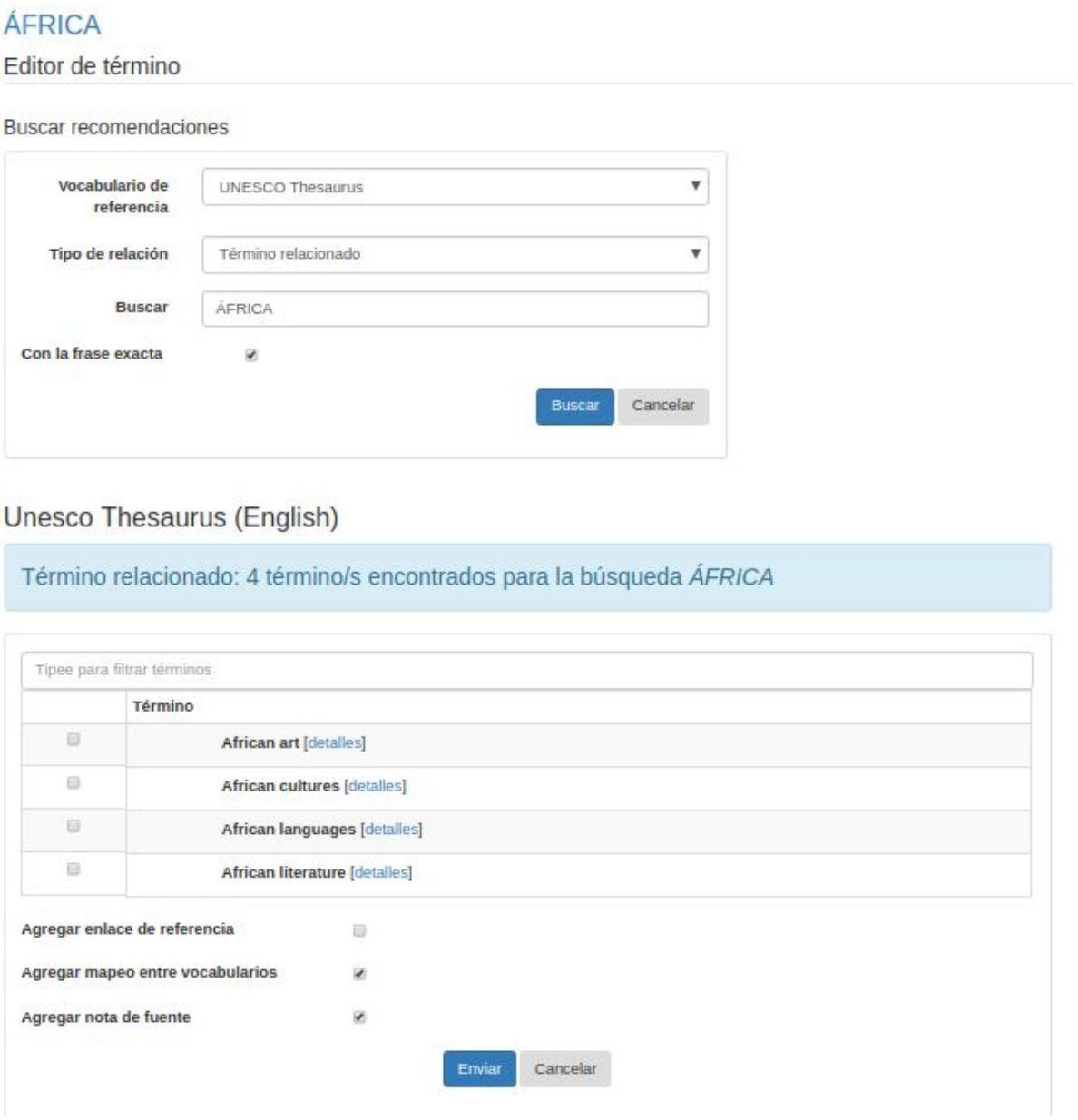

Figura 17. Vista de alta de equivalencia terminológica externa

Interoperabilidad

TemaTres ofrece una API y un endpoint SPARQL que permiten la integración o articulación con otras plataformas de gestión o el desarrollo de servicios derivados basados en la explotación de vocabularios controlados.

El endpoint SPARQL está desarrollado sobre ARC2, una librería que permite trabajar con tripletas RDF almacenándolas en un base de datos en MySQL.

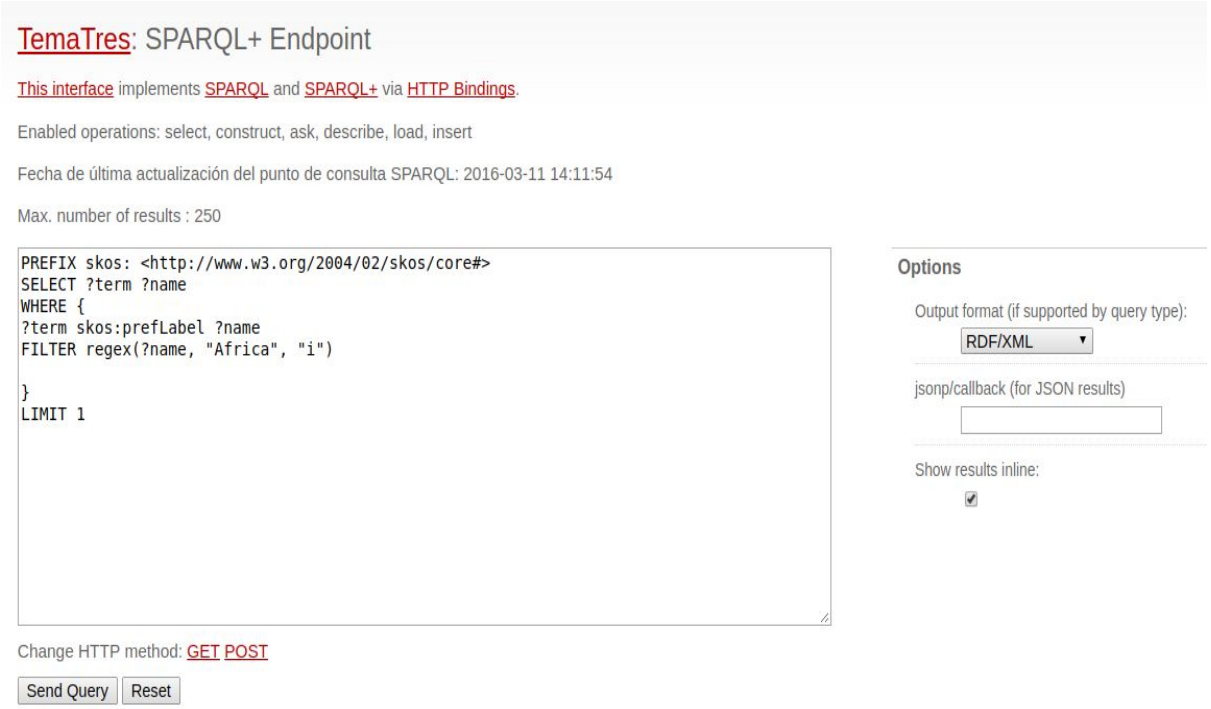

Figura 18. Endpoint SPARQL de TemaTres

En la imagen anterior se hace una consulta SPARQL preguntando por un término que contenga la palabra “África” en la propiedad skos:prefLabel. La respuesta a esa consulta en RDF/XML es la siguiente:

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <!-- query time: 0.0056 sec -->
    <variable name="term"/>
    <variable name="name"/>
  </head>
  <results>
    <result>
      <binding name="term">
        <uri>http://localhost/tematres/?tema=6598</uri>
      </binding>
      <binding name="name">
        <literal xml:lang="es">ABEJA AFRICANIZADA</literal>
      </binding>
    </result>
  </results>
</sparql>
```

La API que ofrece TemaTres es bastante completa y permite recuperar términos bajo diversos criterios a través de consultas HTTP.
Por ejemplo la consulta

`http://localhost/tematres/services.php?task=termsSince&arg=2018-04-05`

permite recuperar datos sobre los términos que se crearon o modificaron desde el 5 de Abril de 2018 en formato XML

```
<?xml version="1.0" encoding="UTF-8"?>
<vocabularyservices>
   <result>
      <term>
         <term_id><![CDATA[10261]]></term_id>
         <code />
         <lang><![CDATA[es]]></lang>
         <string><![CDATA[Controlador]]></string>
         <isMetaTerm><![CDATA[0]]></isMetaTerm>
         <date_create><![CDATA[2018-06-02 07:36:50]]></date_create>
         <date_mod><![CDATA[2018-06-02 07:36:50]]></date_mod>
      </term>
   </result>
   <resume>
      <status><![CDATA[available]]></status>
      <param>
         <task><![CDATA[termsSince]]></task>
         <arg><![CDATA[2017-04-05]]></arg>
      </param>
      <web_service_version><![CDATA[1.6]]></web_service_version>
      <version><![CDATA[TemaTres 3.0]]></version>
   </resume>
</vocabularyservices>
```

Algunas de las consultas que se pueden realizar a la API son:

- Recuperar datos sobre el vocabulario
- Recuperar datos de términos simples por identificador
- Buscar y recuperar términos
- Buscar y recuperar términos usando coincidencia exacta
- Recuperar términos que comienzan con un determinado conjunto de caracteres
- Recuperar términos alternativos un determinado identificador
- Recuperar estructura jerárquica para un determinado término
- Recuperar términos relacionados para un determinado término
- Recupera los últimos términos creados

TemaTres también ofrece la posibilidad de suscribirse a un servicio de sindicación de contenidos a través de RSS.

Este software dispone de diversas herramientas para la generación de reportes y auditorias permitiendo un seguimiento y control de la evolución de un vocabulario controlado. Estos reportes pueden ser utilizados como una herramienta para las rutinas de auditoría y control de calidad que se deseen implementar sobre el vocabulario controlado.

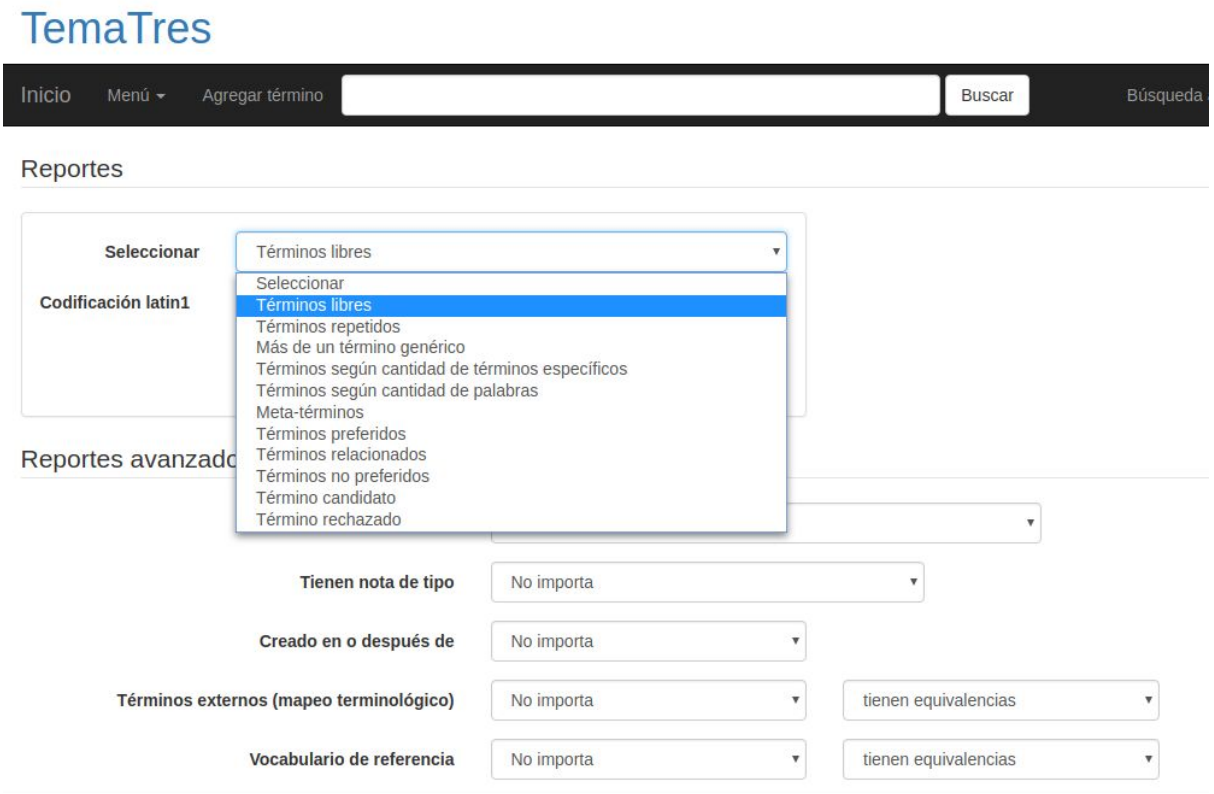

Figura 19. Vista de generación de reportes en TemaTres

Casos de uso de TemaTres como gestor de vocabularios controlados

Se utiliza TemaTres como gestor de vocabularios controlados en el Banco de Vocabularios Jurídicos del SAIJ y el banco semántico del CAICYT.

El Banco de Vocabularios Jurídicos del SAIJ contiene 28 vocabularios entre los que se pueden mencionar:

- Índice de materias Biblioteca Central de la Corte Suprema de la Nación
- Nomenclador terminológico del Ministerio Público Fiscal
- Vocabulario controlado del Consejo de la Magistratura de la Ciudad de Buenos Aires
- Vocabulario controlado de la Fiscalía de Estado de la Provincia de Córdoba
- Tesauro de administración pública
- Vocabulario controlado de la biblioteca del Ministerio de Trabajo

- Vocabulario controlado de la biblioteca del BCRA
- Tesauro del ANSES

Entre los 39 vocabularios que expone el banco semántico del CAICYT podemos mencionar:

- Códigos CONICET de Actividades Industriales
- Vocabulario controlado sobre Biblioclastia
- Códigos CONICET de campo de Aplicación
- Códigos CONICET de disciplinas

Módulo de autoridades basado en CMS

Un sistema de gestión de contenidos, también conocido por sus siglas como CMS (Content Management System), es una aplicación web que permite crear, administrar y publicar de contenido. Estos sistemas cuentan con un módulo de gestión de usuarios que permiten definir usuarios, roles y permisos que van desde administradores del sistema a usuarios sin permisos de edición, o creadores de contenidos.

Entre las principales ventajas de utilizar un CMS podemos enumerar las siguientes:

1. Los CMS suelen ser sistemas muy personalizables, desde el diseño hasta funcionalidad más específicas como configuración dinámica de tipos de contenidos y taxonomías.
2. En el caso de los CMS de código abierto, existen comunidades de programadores que constantemente soluciona fallos, crea nuevos módulos y están en contacto unos con otros para proporcionar el mejor servicio posible a los clientes para los que desarrollan.
3. Actualizaciones de seguridad frecuentes para el caso de una comunidad activa
4. Mejoras en la funcionalidad del sistema, al contar con soporte de plugins o módulos permitiendo añadir nueva funcionalidad en cualquier momento permitiendo a través de nuevos componentes y servicios.

Un CMS cuenta con las características necesarias para implementar un sistema de gestión de vocabularios controlados personalizado que nos permita no sólo estructurar los dichos vocabularios, sino también hacerlos públicos a otros sistemas.
No todos los CMS están orientados a gestionar el mismo tipo de contenido. Algunos CMS tienden a ser diseñados para cumplir diversas expectativas como pueden ser las necesarias por un blog, un foro, una wiki, mientras que otros son diseñados para cumplir con propósitos más generales

Como en este trabajo se requiere desarrollar un sistema de gestión de contenidos muy específicos, se optó por el CMS Drupal ya que este es uno de los más flexibles en comparación con otros CMS como Wordpress o Joomla.

La flexibilidad que posee Drupal se debe a que sus módulos tienen más granularidad ya que sus módulos están pensados para ser reusados desde otros módulos mientras que en Wordpress se piensan para brindar una funcionalidad específica

Por ejemplo para montar un catálogo de autores con Drupal es necesario utilizar distintos módulos, mientras que en CMS como Wordpress o Joomla con un solo módulo es posible tener el mismo resultado aunque con un nivel de personalización definido por el propio módulo y con menor capacidad de adaptación a necesidades más específicas.

Drupal

La arquitectura de Drupal se encuentra conformada por cinco capas que interactúan entre sí. La capa base del sistema está formada por nodos. Los nodos (nodes) son la unidad de información básica en los que Drupal guarda sus contenidos. Un nodo contiene información sobre el autor del contenido, su fecha de creación y posee un título y un cuerpo.

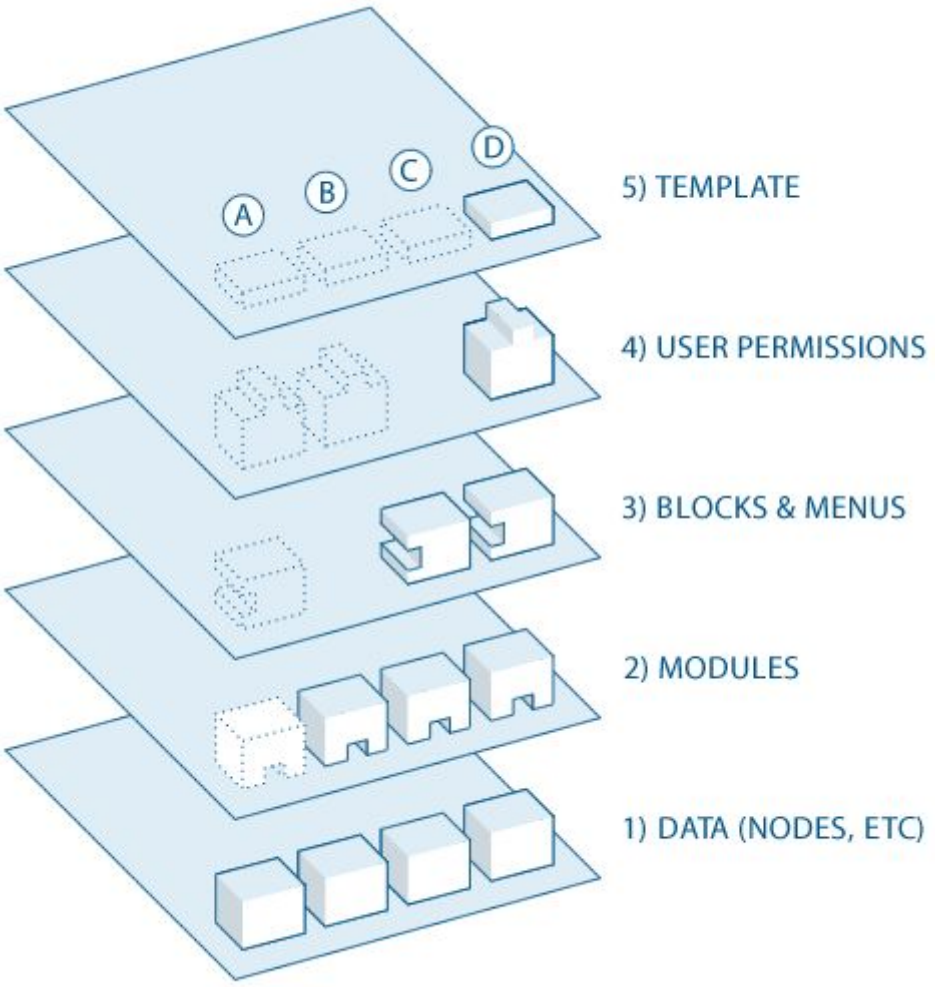

Figura 20. Arquitectura de Drupal («drupal_flow_0.gif (528×562)».)

Cada tipo de nodo se denomina “Content type” (tipo de contenido) y a la información almacenada en los nodos se le denomina “Content”. Existen dos tipos de contenidos básicos que son Article y Basic Page. Los Article representan información que se actualiza con más frecuencia y se categoriza mientras que Basic Page se utiliza para contenido estático que puede vincularse en la barra de navegación principal. El usuario puede habilitar nuevos tipos de contenido así como también implementar uno propio.

La siguiente capa está compuesta por módulos. Los módulos son complementos funcionales que forman parte del núcleo Drupal o han sido creados por miembros de la comunidad Drupal y permiten adaptar cada instancia de acuerdo a las necesidades propias de cada uno.

Los módulos se clasifican en 3 grandes tipos: los “Core Modules”, “Contributed Modules” y los “Custom Modules”.

- Core (núcleo): son los módulos provistos por Drupal al instalarse, por lo cual no requieren ser descargados ni instalados independientemente y pueden ser activados o desactivados desde el back-end. Algunos de ellos fueron contribuciones de la comunidad de Drupal que se incorporaron. Ejemplos: Comments, Node, Taxonomy.
- Contributed (contribuciones): son los módulos que son compartidos para la comunidad de Drupal, están bajo GNU de Licencia Pública (GPL). Se pueden descargar desde la sección de descarga de módulos de drupal.org.
- Custom (personalizados): son los módulos creados por el desarrollador del sitio. Para crearlos se requiere un conocimiento profundo del funcionamiento de Drupal, programación PHP, y la API de Drupal.

En la siguiente capa, se encuentran bloques y menús. Estos albergan y permiten acceder al usuario a la salida generada por los módulos a partir de la información almacenada en los nodos, estructurando y organizando los contenidos en la página web.

La siguiente capa es la de control de usuarios y permisos. Drupal dispone de un registro de usuarios y de roles que permiten especificar qué tareas pueden realizar y a qué contenidos puede acceder cada tipo de usuario definiendo las operaciones que se pueden realizar sobre los elementos provenientes de las capas inferiores.

La última capa de Drupal la forman los temas (themes) y son los principales responsables de la apariencia gráfica o estilo con que se mostrará la información al usuario.

Integración de vocabularios controlados en Drupal

Feeds es un módulo de Drupal desarrollado para importar o agregar contenido en Drupal que trabaja junto a Chaos Tools, un módulo de API con varias herramientas para desarrolladores.

Entre las funciones más importantes de Feeds podemos mencionar

- importar fuentes RSS, fuentes Atom, archivos OPML o archivos CSV, XML o HTML
- generación de usuarios, nodos, términos o simples registros de bases de datos
- mapeo de propiedades o configuraciones con archivos.
- aplicar múltiples configuraciones de forma simultáneas organizadas en entidades llamadas "Importadores" (importers)
- Importación periódica en cron
- exportar estilos de vistas

Interoperabilidad en Drupal

Drupal dispone de varios módulos dedicados para interoperar con otros sistemas, como módulos para definir una API REST, exponer información en formato RDFa o sindicar contenido a través de RSS. Habiendo analizado, en el capítulo 3, las ventajas que conlleva seguir los principios de Linked Data en este trabajo interesa poder exponer las autoridades a través de un endpoint SPARQL.

Existe un módulo llamado ARC2 store que permite almacenar en una base de datos un conjunto de entidades y recuperarlas por medio de SPARQL. Para eso Arc2 store funciona junto con el módulo RDF indexer cuya función es la de mapear los tipos de contenido definidos en Drupal a RDF e indexarlos para que estos puedan ser expuestos a otros sistemas.

Casos de uso de Drupal como gestor de vocabularios controlados

Alguno de los casos donde Drupal es utilizado como gestor de vocabularios controlados es Bartoc y el Tesauro de la Unión Europea EuroVoc. Ambos casos fueron presentados en el capítulo 2 y se corresponden a dos instituciones que manejan grandes cantidades de vocabularios controlados, sobre todo el caso de EuroVoc que, entre sus instituciones más importantes, es utilizado por el Parlamento Europeo, la Oficina de Publicaciones, administraciones nacionales europeas y usuarios privados tanto de los países miembros de la UE como de países terceros.

Capítulo 6 - Análisis y propuesta de solución

Introducción

El repositorio institucional de la UNLP, SEDICI, se encuentra desarrollado en DSpace, software que permite gestionar autoridades de forma externa. En este caso SEDICI utiliza el concepto de autoridades para representar personas e instituciones, en una aplicación llamada CelsiusDL. También gestiona tesauros y sistemas de clasificación.
Actualmente existen nuevas alternativas que permiten gestionar autoridades, y tras lo analizado en el capítulo 5, se decidió migrar los vocabularios controlados almacenados en CelsiusDL a un sistema personalizado construido sobre Drupal. En este capítulo se explicará en qué estado se encuentra actualmente CelsiusDL, las razones por las cuales se eligió Drupal como sistema de gestión de autoridades y cuál fue el plan de migración a seguir.
A grandes rasgos el plan de migración aplicado fue el siguiente:

1. Relevar modelo de datos y funcionamiento de CelsiusDL
2. Definir requerimientos para elección de herramienta
3. Desarrollar módulos de Drupal que permitan
    a. Modelar entidades y sus correspondientes mapeos a RDF
    b. Exponer entidades importadas a través de SPARQL
4. Normalizar entidades en CelsiusDL y exportarlas.
5. Crear entidades en Drupal a partir de las autoridades exportadas anteriormente
6. Desarrollar módulo en DSpace que permita consumir los datos expuestos en el endpoint SPARQL
7. Actualizar referencias a las autoridades en CelsiusDL por las autoridades en Drupal

Estado de autoridades en SEDICI

Las autoridades usadas en el repositorio SEDICI, se encuentran en un base de datos MySQL, externa a DSpace, que se gestiona a través de una aplicación independiente llamada CelsiusDL. Esta aplicación, desarrollada en Java con tecnologías antiguas fue parte de la implementación original del repositorio en 2003 pero debe ser abandonada cuanto antes ya que ha caducado tanto funcional como tecnológicamente. Otro de los problemas que evidencia esta aplicación es la falta de mantenimiento y actualizaciones que, sumado a los continuos cambios que se presentan en SEDICI, ha provocado no solo problemas de diseño sino también redundancia de datos.
Los metadatos controlados en SEDICI obtienen las autoridades de las tablas *tesaruro_termino*, *personas* y *jerarquias_termino* de la base de datos de CelsiusDL.

Tabla	Entidad representada
jerarquias_termino	Instituciones de Argentina Grado a alcanzar
tesaruro_termino	Tesauro de UNESCO Descriptores SeDiCI Materias Tesauro de Eurovoc Tesauro DeCS
personas	Personas
jerarquias_relaciones	Tabla que representa la relación entre términos de la tabla *jerarquias_termino*
tesauros_relaciones	Tabla que representa la relación entre términos de la tabla *tesauro_relaciones*

Tabla 2. Entidades almacenadas en sus respectivas tablas en CelsiusDL

Como DSpace no gestiona autores, la filiación de estos no se controla en SEDICI, aunque si se almacena esta relación en CelsiusDL. Las filiaciones de las personas están representadas en las tablas *paises*, *instituciones*, *dependencias* y *unidades* relacionándose entre sí, formando una jerarquía donde una unidad se corresponde con una dependencia que a su vez se corresponde con una institución que pertenece a un país. Los identificadores de estas 4 entidades se utilizan para representar la filiación de un autor. Por ende la tabla *personas* además de las columnas que representan la información básica de una persona (*nombre*, *apellido* e *email*) tiene 4 columnas destinadas a representar la filiación siendo estas *id_pais*, *id_institucion*, *id_dependencia* y *id_unidad*.

Como se pudo apreciar anteriormente, no solo existe información sobre organizaciones e instituciones en las 4 tablas mencionadas anteriormente, sino que también existen algunas de estas entidades en una la tabla *jerarquía_terminos*. Este error de diseño conlleva a problemas como la duplicación de información, inconsistencia de datos y un mantenimiento más engorroso de las entidad almacenadas.

Interconexión actual con DSpace

Para acceder a las autoridades en CelsiusDL desde SEDICI se han implementado tres clases que extienden *org.dspace.content.authority.ChoiceAuthority.java* que es la clase que provee DSpace como interfaz del plugin de choice authority.
Estas 3 clases son:

- ar.edu.unlp.sedici.dspace.authority.SeDiCI2003Authors.java
- ar.edu.unlp.sedici.dspace.authority.SeDiCI2003Jerarquia.java
- ar.edu.unlp.sedici.dspace.authority.SeDiCI2003Tesauro.java

Básicamente estas clases se encargan de adaptar los datos traídos de la fuente de datos externa al modelo esperado por DSpace y su módulo de autoridades. En el siguiente cuadro se puede observar sobre qué metadatos se aplica el control de autoridades y que clase se encarga de ellos.

Clase	**Metadatos**
ar.edu.unlp.sedici.dspace.authority.SeDiCI2003Authors	sedici.creator.person sedici.contributor.codirector sedici.contributor.director sedici.contributor.compiler sedici.contributor.inscriber sedici.creator.interprete sedici.contributor.juror
ar.edu.unlp.sedici.dspace.authority.SeDiCI2003Jerarquia	sedici.creator.corporate sedici.institucionDesarrollo thesis.degree.name thesis.degree.grantor mods.originInfo.place
ar.edu.unlp.sedici.dspace.authority.SeDiCI2003Tesauro	sedici.subject.eurovoc sedici.subject.materias sedici.subject.decs sedici.subject.descriptores dc.coverage.spatial

Tabla 3. Clases que controlan metadatos en SEDICI

Problemas de la solución actual

Además de la antigüedad que presenta CelsiusDL y de la duplicación de información explicado anteriormente, este sistema carece de chequeos de validez lo que llevó a problemas de integridad de datos. CelsiusDL presenta otro problema que es la gran cantidad de información almacenada que no es utilizada. Esto se debe a que con el transcurso del tiempo se han tomado decisiones como por ejemplo almacenar todos los autores de todas la obras o todas las instituciones relacionadas a una publicación. Esto no necesariamente debe ser así, ya que existen casos donde la autoridad generada no será consultada nunca más, o no es información que SEDICI considere lo suficientemente importante como para crear una autoridad. Cabe recordar que una de las funciones que deben cumplir el uso de autoridades es la de identificar conceptos y realizar las búsquedas de los mismos de la forma más eficiente posible por ende mientras mayor sea el volumen de datos, más dificultosa será esta tarea. CelsiusDL tampoco contempla la definición de roles, lo cual es importante ya que no todos los usuarios que administran las autoridades en SEDICI tienen las mismas funciones.

Elección de una herramientas para gestión de autoridades

Como se planteó en el capítulo 1, el objetivo de este trabajo es implementar un sistema de gestión de autoridades que tenga la posibilidad de modelar autores, sus filiaciones y algunos tesauros y taxonomías. Este sistema debe poder modelar al menos dos perfiles de usuarios distintos a los que llamaremos administrador y editor.

El administrador debe tener los privilegios necesarios para administrar el contenido del sistema junto con su configuración. Los usuarios con perfil de editor deben ser capaces de cargar información sin la necesidad de conocer el modelo subyacente ni cómo se implementa la interoperabilidad entre este nuevo sistema y SEDICI.

Otro de los aspectos a tener en cuenta es la capacidad de migrar los datos almacenados en el viejo sistema al nuevo, sin modificar la forma en que el repositorio gestiona dicha información.

Tras haber realizado un análisis de las distintas alternativas que nos permiten desarrollar un sistema de gestión de vocabularios controlados se optó por escoger el CMS Drupal, detallando a continuación las razones de la elección de dicha herramienta así como las razones para descartar TemaTres y VocBench.

Se descarta el uso de TemaTres para implementar este sistema ya que la forma en que la herramienta modela los vocabularios controlados y los estados que el mismo puede atravesar, no aplican en el contexto de una base de datos de autoridades. El sistema debe disponer un modelo de datos flexible que permita representar entidades de todo tipo y que

permita una mutación de los mismos en el tiempo sin embargo en TemaTres los vocabularios controlados son gestionados como un tesauro y posee estados (candidatos, aceptados o rechazados) que no interesan en el contexto de nuestra aplicación. Por último esta herramienta tampoco dispone de algún módulo que permita importar los datos del sistema de gestión de autoridades viejo, lo cual termina descartando definitivamente esta propuesta.

El caso de Vocbench es distinto al de TemaTres. Esta herramienta está diseñada para gestionar ontologías OWL, tesauros SKOS (XL) para satisfacer las necesidades de la web semántica lo que hace que los usuarios tengan conocimientos sobre OWL, SKOS(XL) y RDF dificultando la tarea de los usuarios con rol de editor.

El propósito y las funcionalidades de VocBench son demasiado específicas para las necesidades del sistema de gestión de vocabulario necesario en SEDICI. Otro de los inconvenientes que tiene esta herramienta yace en la importación de contenido. VocBench permite importar ontologías y tesauros, pero no permite generar instancias de la información a migrar, al menos no sin implementar algo desde cero.

La elección de un CMS como base para el sistema de gestión de vocabularios controlados se basa en una herramienta pensada de forma genérica, más personalizable que las alternativas mencionadas anteriormente. La elección de Drupal como herramienta para implementar un gestor de autoridades se debe varios motivos, a saber:

- capacidad para definir nodos con estructura dinámica
- posibilidad de reutilizar módulos que faciliten el modelado,
- herramientas de importación de contenidos con sus relaciones
- herramientas de indexación
- La posibilidad que tiene Drupal de poder exponer su contenido a otros sistemas a través de SPARQL
- posibilidad de definir roles para administrador y editor terminaron siendo fundamentales en esta elección, entre otras funcionalidades.
- Casos de éxito como Bartoc y Tesauro de la Unión Europea EuroVoc

Solución propuesta

La solución propuesta basada en Drupal, contempla el desarrollo de un módulo personalizado (o como llamamos en el capítulo 6, módulo custom), llamado AuthVoc, donde las autoridades se guardan como nodos en una instalación de Drupal 7 de acuerdo a un modelo definido por content-types, donde el identificador de cada autoridad se corresponde con la URI del recurso, siguiendo uno de los principios de Linked Data. Las propiedades de estos content-types personalizados se mapean en tripletas y se almacenan en una base de

datos RDF para luego ser ser expuestos a través de un servicio de búsqueda SPARQL. Este endpoint permite realizar búsquedas en formato SPARQL y recuperar los datos en formato JSON, CSV y XML. También está contemplado el desarrollo de submódulos que permitan importar desde un CSV la información normalizada del sistema de autoridades CelsiusDL. Para poder recuperar las autoridades definidas en este nuevo sistema, será necesario extender el módulo de autoridades en DSpace para que asocie los metadatos a controlar con la nueva fuente de autoridades.

En esta primera versión del sistema solo se migrarán las personas y sus correspondientes filiaciones, dejando para un futuro la incorporación de tesauros, materias y grados alcanzados.

La siguiente imagen ilustra cómo el framework de Choices de DSpace realiza consultas SPARQL al módulo AuthVoc de Drupal y cómo este interactúa con los submódulos que lo componen.

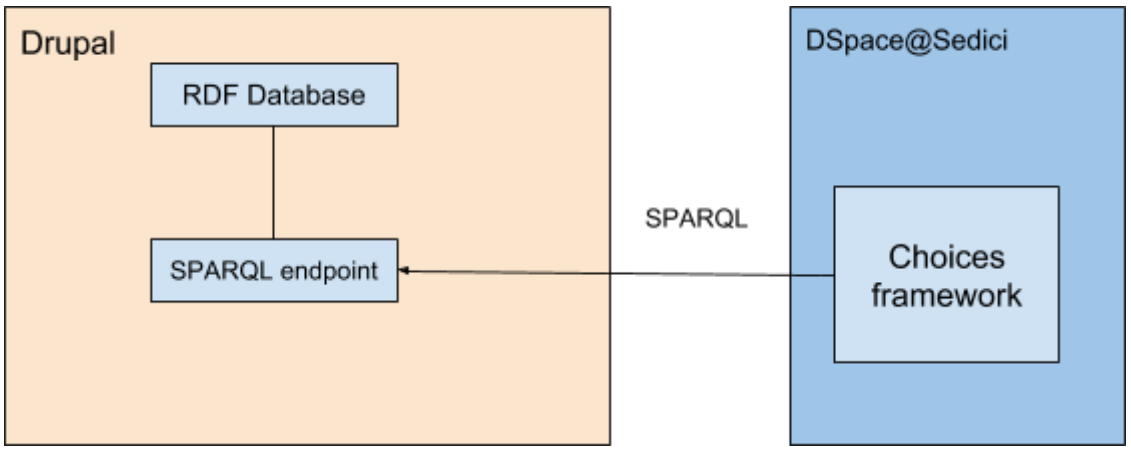

Figura 21. Diagrama de solución propuesta

Capítulo 7 - Desarrollo y migración

Introducción

En el capítulo 6 se expusieron los problemas que tiene el software CelsiusDL y se propuso migrar las autoridades gestionadas por dichos programa a un sistema basado en Drupal. Drupal permite agregar funcionalidad a través del desarrollo de módulos que pueden ser desarrollados por una comunidad ya consolidada o desarrollados acorde a las necesidades personales de cada uno. Para agregar la funcionalidad deseada se implementó un módulo llamada AuthVoc, que tiene como principal objetivo estructurar las autoridades gestionadas anteriormente en CelsiusDL, y exponerlas a través de SPARQL.

En este capítulo se explicarán los módulos que conforman este nuevo sistema, sus funcionalidades respectivas y como estos se relacionan entre sí.

Desarrollo de AuthVoc

El módulo AuthVoc debe poder representar las entidades a controlar, exponerlas a otros sistemas y permitir importar las autoridades del sistema a migrar. Para cada una de estas funcionalidades se desarrollaron los siguientes submódulos:

- AuthCTypes: modela los tipos de contenido definidos para Persona e Institución y mapea sus propiedades a RDF
- Auth SPARQL: indexa y expone a través de un endpoint SPARQL las instituciones y las personas.
- Auth Importer: compuesta por módulos que crean nodos a partir de archivos en formato CSV originados a partir de los datos a migrar del sistema CelsiusDL.

La siguiente imagen ilustra cómo AuthVoc, y los submódulos que lo componen, se construyen sobre un conjunto de módulos Contrib reutilizando así la funcionalidad ya implementada por la comunidad de desarrolladores de Drupal.

Módulo AuthCTypes

El módulo AuthCTypes define un modelo que soluciona uno de los problemas que existía en CelsiusDL donde una persona puede tener sólo una filiación. El modelo propuesto permite relacionar a una persona con una o más instituciones, registrando también la fecha de inicio, fecha de fin y rol que cumple en dicha institución.

El diagrama de clases propuesto es el siguiente

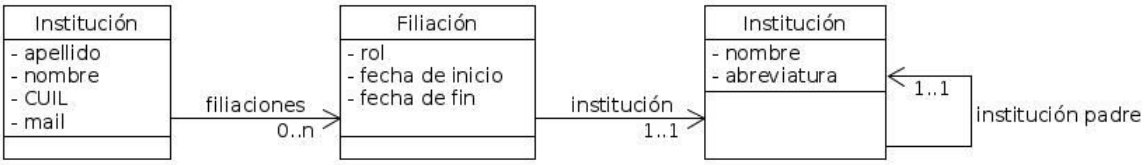

Diagrama X. Entidades representadas en AuthCTypes

Para poder exponer estas entidades a través de SPARQL fue necesario mapear los campos a publicar con un conjunto de ontologías, como se muestra en las siguientes tablas

Persona	
RDF Types: foaf:Person	
Campo	**Predicado RDF**
apellido	foaf:familyName
nombre	foaf:givenName
CUIL	dc:identifier
mail	foaf:mbox
filiaciones	cerif:linksToOrganisationUnit

Tabla 4. Mapeo de atributos de entidad Persona con predicado RDF

Filiación	
RDF Types: sioc:Post, sioc:Comment	
Campos	**RDF PREDICATES**
institución	foaf:Organization
rol	cerif:role
fecha de inicio	cerif:startDate
fecha de fin	cerif:endDate

Tabla 5. Mapeo de atributos de entidad Filiación con predicado RDF

Institución	
RDF Types: sioc:Item, foaf:Organization	
Campo	**Predicado RDF**
nombre	foaf:name
abreviatura	sioc:id
institución padre	foaf:Organization

Tabla 6. Mapeo de atributos de entidad Institución con predicado RDF

Módulo AuthSPARQL

El módulo AuthSPARQL fue desarrollado para exponer los campos de las entidades Persona e Institución y poder ser accedidos desde DSpace a través de un endpoint desarrollado sobre ARC2Store, de la misma forma que lo hace el software TemaTres.

Para esto fue necesario utilizar los módulos contrib RDF indexer y SearchAPI. El módulo RDF indexer tienen como función indexar recursos en forma de tripletas almacenados en una base de datos RDF, en este caso ARC2, mientras que Search API es un módulo que permite realizar búsquedas sobre cualquier tipo de entidad de Drupal.

En la figura 22 se puede apreciar como es el endpoint SPARQL provisto por ARC2.

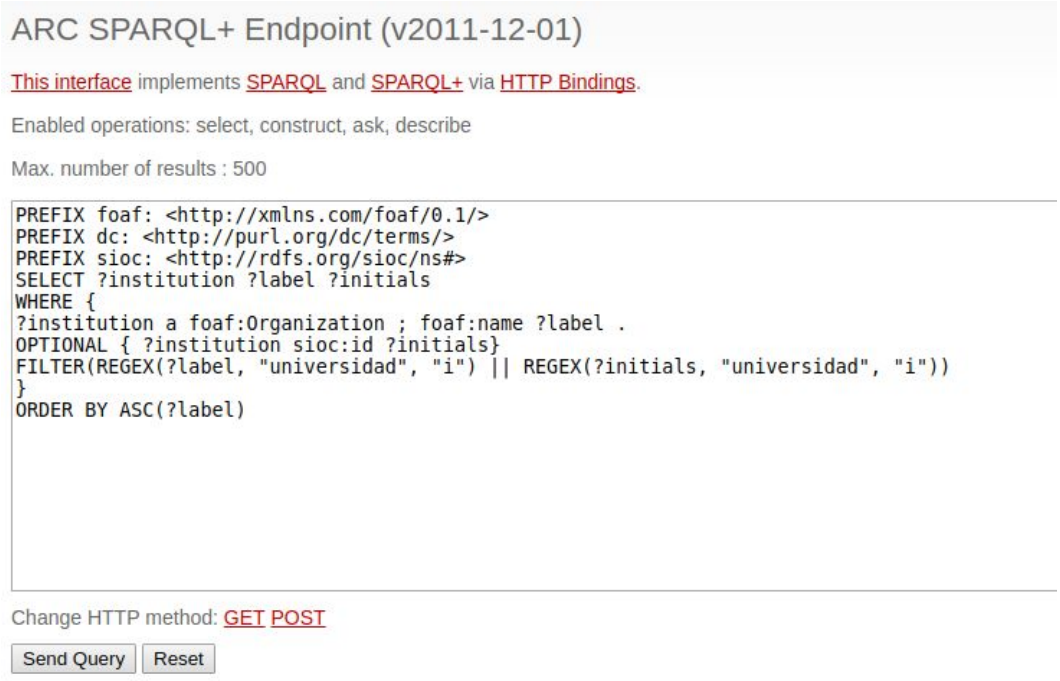

Figura 22. Vista de endpoint SPARQL

De acuerdo al esquema propuesto en el módulo AuthCTypes, es posible consultar por instituciones y personas, junto con sus filiaciones. Por ejemplo la siguiente consulta retorna un texto con el uri del recurso, el nombre de la institución y su abreviatura en caso de que tuviese:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX sioc: <http://rdfs.org/sioc/ns#>
SELECT ?institution ?label ?initials
WHERE {
?institution a foaf:Organization ; foaf:name ?label .
OPTIONAL { ?institution sioc:id ?initials}
FILTER(REGEX(?label, "universidad", "i") || REGEX(?initials, "universidad", "i"))
}
ORDER BY ASC(?label)
```

La respuesta en formato RDF/XML es la siguiente:

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <!-- query time: 0.0307 sec -->
    <variable name="institution"/>
    <variable name="label"/>
    <variable name="initials"/>
  </head>
  <results>
    <result>
      <binding name="institution">
        <uri>http://localhost/auth-voc/?q=node/46624</uri>
      </binding>
      <binding name="label">
        <literal>Universidad Nacional de La Plata</literal>
      </binding>
      <binding name="initials">
        <literal>UNLP</literal>
      </binding>
    </result>
  </results>
</sparql>
```

La siguiente consulta devuelve un grafo RDF, ya que se realiza un CONSTRUCT en vez de una SELECT. Se utiliza esta consulta ya que una persona puede tener uno o más instituciones asociadas, y es más fácil representar esta relación en un grafo que mediante un String.

```
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  foaf: <http://xmlns.com/foaf/0.1/>
PREFIX  dc:   <http://purl.org/dc/terms/>
PREFIX  cerif: <http://spi-fm.uca.es/neologism/cerif/1.3#>
PREFIX  sioc: <http://rdfs.org/sioc/ns#>

CONSTRUCT
  { ?person rdf:type foaf:Person .
    ?person foaf:givenName ?name .
    ?person foaf:mbox ?mail .
    ?person foaf:surame ?surname .
    ?person cerif:linksToOrganisationUnit ?link .
    ?link cerif:startDate ?inicio .
    ?link cerif:endDate ?fin .
    ?link foaf:Organization ?org .
    ?org foaf:name ?affiliation .
    ?org sioc:id ?id .}
WHERE
  { ?person  rdf:type         foaf:Person ;
             foaf:givenName   ?name ;
             foaf:surname  ?surname .
    OPTIONAL
      { ?person  foaf:mbox  ?mail . }
    OPTIONAL
      { ?person  cerif:linksToOrganisationUnit  ?link .
        ?link    cerif:startDate       ?inicio ;
                 cerif:endDate         ?fin ;
                 foaf:Organization     ?org .
        ?org     foaf:name             ?affiliation ;
                 sioc:id               ?id .
      }
    FILTER ( ( regex(?name, "martinez", "i") || regex(?surname, "martinez", "i") )
|| regex(?id, "martinez", "i") )
  }
ORDER BY ?surname ?link
OFFSET  0
LIMIT   1
```

La respuesta en RDF/XML a la consulta anterior sería:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:ns0="http://xmlns.com/foaf/0.1/"
  xmlns:ns1="http://spi-fm.uca.es/neologism/cerif/1.3#"
  xmlns:ns2="http://rdfs.org/sioc/ns#">

  <rdf:Description rdf:about="http://localhost/auth-voc/?q=node/25084">
    <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
    <ns0:givenName>Nadina Martinez</ns0:givenName>
    <ns0:surame>Carod</ns0:surame>
    <ns1:linksToOrganisationUnit
         rdf:resource="http://localhost/auth-voc/?q=field_collection_item/7947"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://localhost/auth-voc/?q=field_collection_item/7947">
    <ns1:startDate></ns1:startDate>
    <ns1:endDate></ns1:endDate>
    <ns0:Organization rdf:resource="http://localhost/auth-voc/?q=node/46701"/>
  </rdf:Description>

  <rdf:Description rdf:about="http://localhost/auth-voc/?q=node/46701">
    <ns0:name>Universidad Nacional del Comahue</ns0:name>
    <ns2:id>UNCo</ns2:id>
  </rdf:Description>
</rdf:RDF>
```

Proceso de migración

Las etapas que se llevaron a cabo para migrar las autoridades desde CelsiusDL a Drupal fueron:

1. Normalización de información almacenada en la base de datos de CelsiusDL
    a. Unificación de instituciones.
    b. Corrección de errores de sintaxis.
2. Eliminación de autoridades innecesarias
3. Exportación de autoridades a formato CSV
4. Importación de autoridades en AuthVoc

1. Normalización

Para unificar las instituciones se creó una tabla intermedia, llamada *export_instituciones,* que fue utilizada para unificar la información de estas entidades, donde para cada institución

se almacenan los identificadores de otras tablas que hagan referencia a dicha entidad. De esta forma en un tupla se centraliza toda las referencias correspondiente a una entidad.
A continuación se muestra un diagrama con las tablas involucradas en la migración. Se puede observar cómo se relacionan las tablas *instituciones*, *dependencias* y *unidades* para representar una entidad que también puede estar representada en la tabla *jerarquia_termino*.

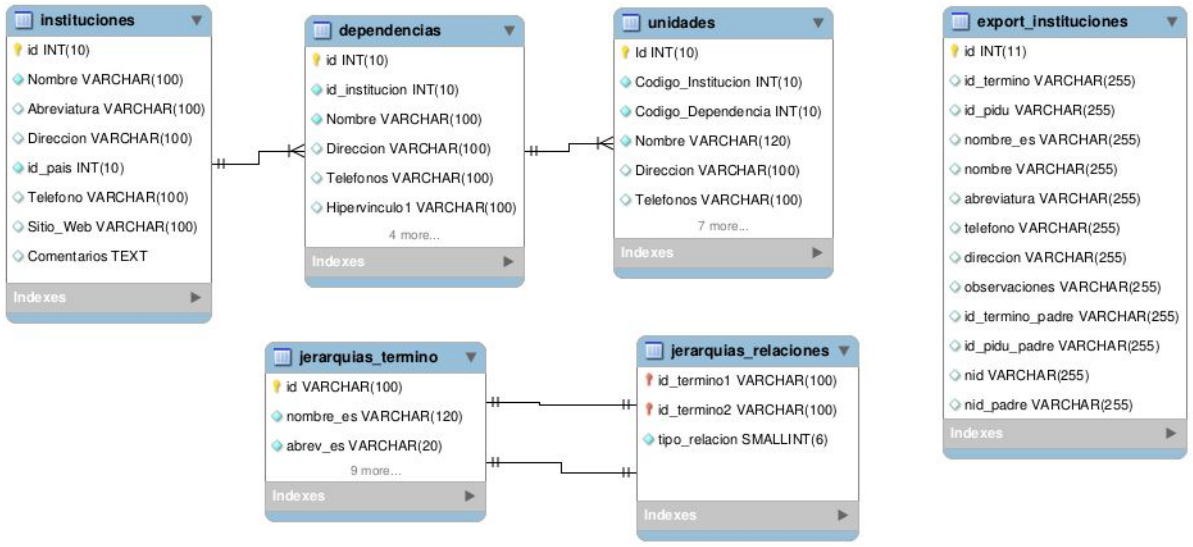

Figura 23. Tablas que representan instituciones en CelsiusDL

Mientras se realizaba la unificación de instituciones se fueron corrigiendo los errores encontrados, siendo algunos de ellos errores de sintaxis y abreviaturas escritas en el nombre en vez de encontrarse en la columna dedicada a esta información. Una vez unificado las instituciones se actualizaron las filiaciones de los autores, ya que para cada institución se generó un identificador nuevo.

2. Eliminación de autoridades innecesarias

Como se mencionó en el capítulo 6, existe un gran número de autoridades en CelsiusDL que dificultan el mantenimiento de la base de datos y dificultan las tareas de control de autoridades. No todas las autoridades tienen la misma importancia, ya sea por lo que representan o por la cantidad de veces a la que se hace referencia a dicha entidad. Es por eso que una vez normalizado todas las autoridades se procedió a hacer un borrado de aquellos recursos que no se consideren importantes para SEDICI.

Para definir la importancia de una una institución se tuvo en cuenta:

1. País de origen de la institución
2. Si forma parte de la UNLP u otras instituciones argentinas de importancia como CIC o CONICET

3. Cantidad de publicaciones

Para definir el peso de un autor se hizo hincapié en:

1. Institución a la que pertenece
2. Cantidad de publicaciones
3. Fecha de su última publicación

Para definir el grado de importancia que tiene una autoridad fue necesario cruzar la información disponible en la base de datos en MySQL de CelsiusDL, como identificadores de personas e identificadores de sus filiaciones, con información de los autores en SEDICI, almacenados en una base de datos en PostgreSQL, de donde se puede obtener cantidad de publicaciones por autor y sus fechas. Para esto se crearon dos tablas intermedias que agrupan toda la esta información, como se puede apreciar en la figura 24.

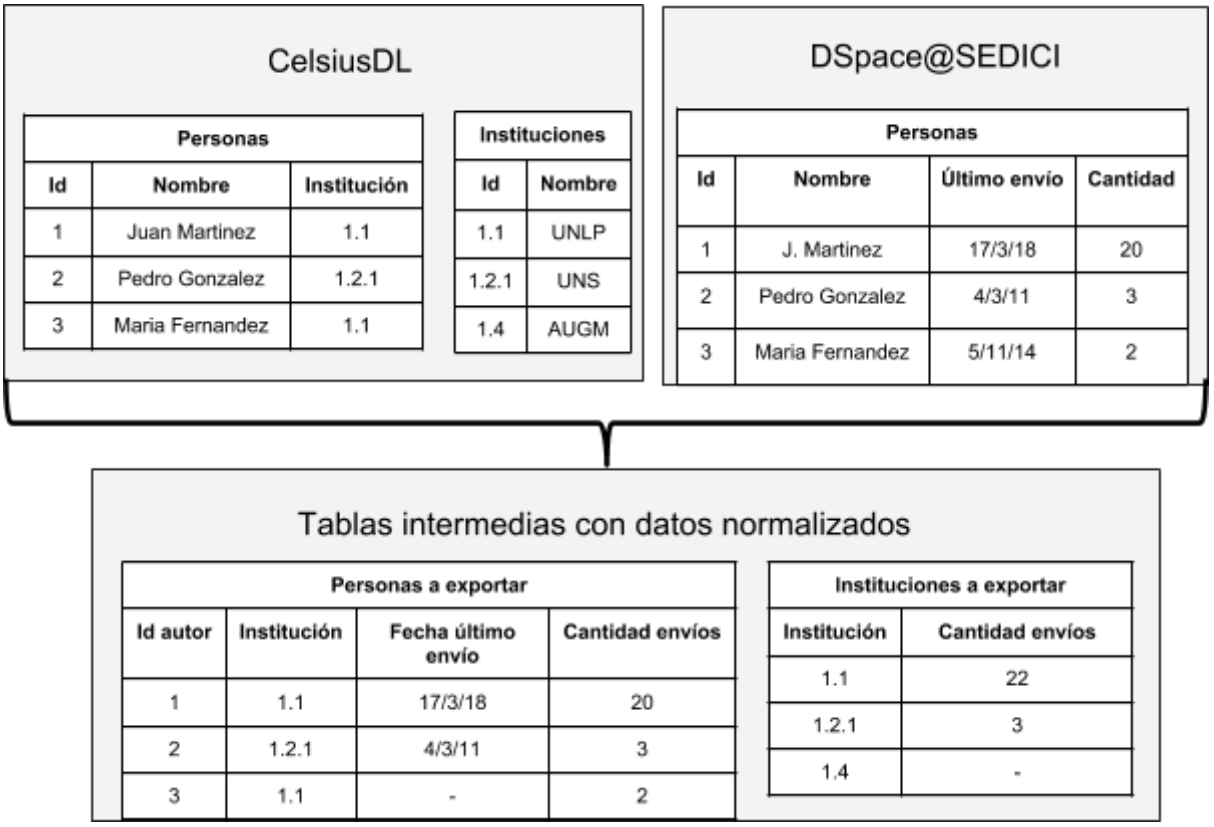

Figura 24. Ejemplo de tablas intermedias generadas para eliminar autoridades innecesarias

Una vez obtenida esta información y habiendo cruzado estos resultados con las autoridades en CelsiusDL se procedió a eliminar entidades que cumpliesen con alguna de las siguiente condiciones:

- Personas que tenían sólo 1 item y no pertenecían a la UNLP
- Autores que no son de Argentina y que tenían 2 items o menos y cuyo último haya sido 2017 o antes.

- Autores que no son de argentina y que tenían participación en 3 o menos ítems, cuyo último envío fuese en 2013 o antes.
- Autores no argentinos que pertenecen a instituciones que hayan hecho 10 o menos publicaciones.
- Instituciones argentinas que no pertenecían a la UNLP o al CONICET con menos de 10 publicaciones
- Autores pertenecientes a la UNLP con 1 solo item cuyo envío tenga fecha anterior al 2011.

De esta forma se logró reducir el número autoridades de 46900 autores a 18600 y de 1536 instituciones 1428.

3. Exportación de autoridades a formato CSV

Dentro de la información a exportar, no solo es necesario tener en cuenta los datos propios de cada entidad (como nombre y apellido en el caso de las personas) sino también las claves de las autoridades en CelsiusDL, ya que más tarde será necesario actualizar las referencias que se hagan en DSpace a estas entidades por las nuevas autoridades almacenadas en el nuevo sistema.

CSV ejemplo de personas a importar:

nid	old_id	apellido	nombre	dni	cargo	filiacion
48368	12649	González	Carlos		Becario	48024
48369	12650	Mármol	Juan			48024

CSV ejemplo de instituciones a exportar

nid	nid padre	nombre	abrev	id termino	id termino padre	id pidu	id pidu padre
46940		Universidad Nacional de La Plata	UNLP	1.1	1	1	1
46941		Comisión Nacional de Energía Atómica	CNEA	1.15	1	5	1

4. Importación de autoridades en AuthVoc

Para importar los archivos CSVs generados en la exportación de autoridades se desarrolló un módulo personalizado, llamado AuthImporter, basado en el módulo Feeds. Feeds permite definir una entidad, llamada importer, cuya función es mapear los campos en un archivo CSV generado con los campos de un tipo de contenido, de esta forma es posible crear un nuevo nodo o actualizar uno existente. Para crear las entidades a partir de los archivos CSVs se crearon tres importers, un para importar instituciones, otro para importar personas y un último importer para generar la relación entre personas y las instituciones, formando así la filiación del autor.
A continuación se muestra la vista del importer donde se mapean los campos de un CSV que contiene autores, con los campos de un tipo de contenido Persona

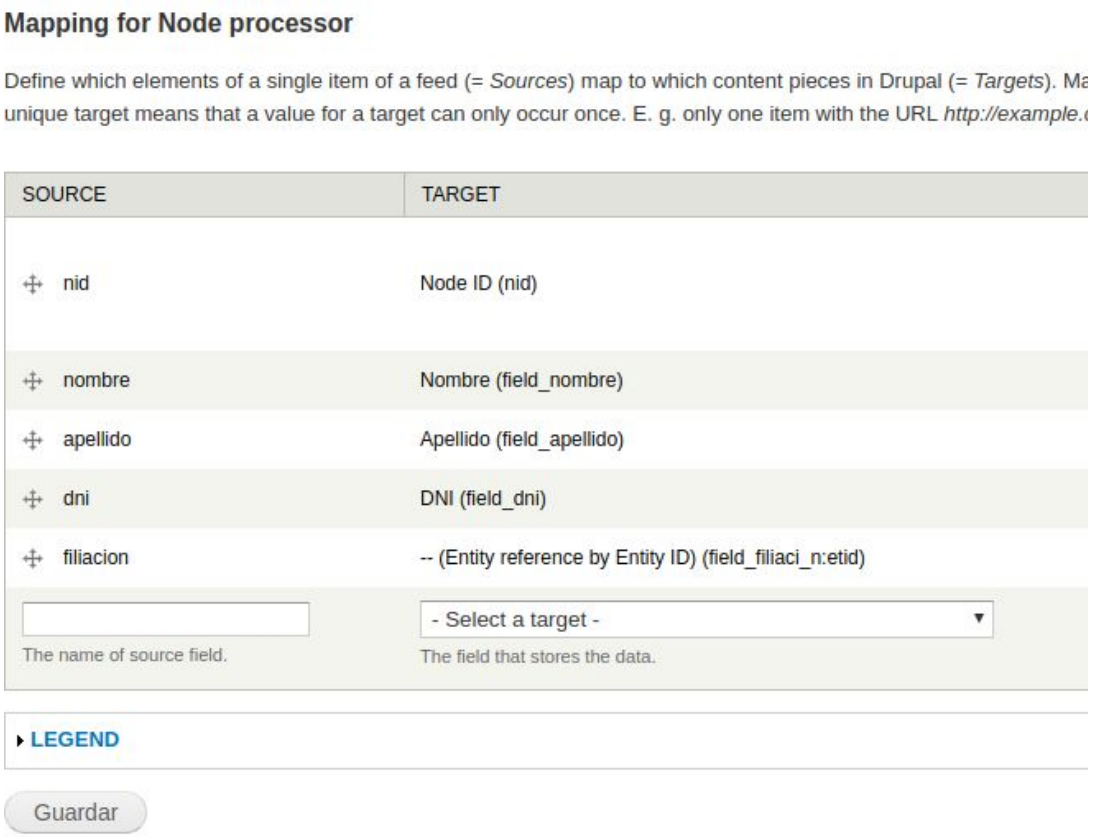

Figura 25. Mapeo de atributos de entidad Persona con campos de archivo CSV

Como explicamos en el capítulo 4, DSpace asocia, en su base de datos, metadatos controlados con una clave de autoridad. Al cambiar la fuente de autoridades, fue necesario reemplazar el valor del campo authority_key que referencia a una autoridad en CelsiusDL por la clave de la autoridad correspondiente en la instalación de Drupal.

Drupal permite generar vistas personalizadas que pueden ser exportadas en archivos CSVs, por lo que fue posible crear un archivo con entradas para cada una de las

autoridades migradas y sus claves de autoridad correspondiente, tanto para CelsiusDL (ya que las mismas formaron parte de la migración) como las nuevas claves en Drupal.

export instituciones

Nid	id_termino
48024	1.1.13.3
48040	1.1.23.7
48056	1.1.23.9.1
48025	1.1.15.1
48041	1.1.23.8
48057	1.1.3.14.5
48026	1.1.15.2
48042	1.1.23.9
48058	1.1.3.14.8
48027	1.1.2.4

1 2 3 4 5 6 7 8 9 … siguiente › última »

Figura 26. Vista de nuevos identificadores de institución con sus correspondientes identificadores en CelsiusDL

La figura 26 muestra cómo es una vista con las identificadores de los nodos en Drupal y su correspondiente identificador en DSpace. Luego de haber exportado la vista anterior a un archivo CSV, se edita el mismo para generar múltiples consultas SQL que actualicen las claves de autoridad en la base de datos de DSpace en SEDICI. Si nos basamos en la figura 26, la primera línea del csv sería "`48024, 1.1.13.3`" y la consulta generada en base a esa línea debería actualizar todos las claves de autoridad con valor "`1.1.13.3`" con la uri del nodo "`48024`" que sería por ejemplo `http://[URL]/node/48024`

Implementación de conector en DSpace

En el capítulo 4 se mencionó al módulo de autoridades de DSpace y los beneficios que conlleva su uso. Para eso DSpace permite extender un conjunto de clases e interfaces que se encargarán de recuperar las autoridades definidas externamente asociando estas los resultados obtenidos con un conjunto de metadatos a controlar.
En base a las clases provistas por DSpace para extender la funcionalidad deseada se procedió a la implementación de las siguientes clases:

- SPARQLAuthorityProvider.java: Clase Abstracta que implementa ChoiceAuthority que sirve de base para definir fuentes de autoridades específicas.

- AdvancedSPARQLAuthorityProvider.java: Clase que contiene la funcionalidad para procesar consultas SPARQL con CONSTRUCT
- SimpleSPARQLAuthorityProvider.java: Clase que contiene la funcionalidad para procesar consultas SPARQL con SELECT
- AuthorAuthority.java e InstitutionAuthority.java: Clases concretas que definen las consultas SPARQL que se enviarán al endpoint y procesan los resultados de la respuesta, reemplazando a las clases SeDiCI2003Authors y SeDiCI2003Institutions mencionadas en el capítulo 5, controlando así al mismo conjunto de metadatos definido en SEDICI.

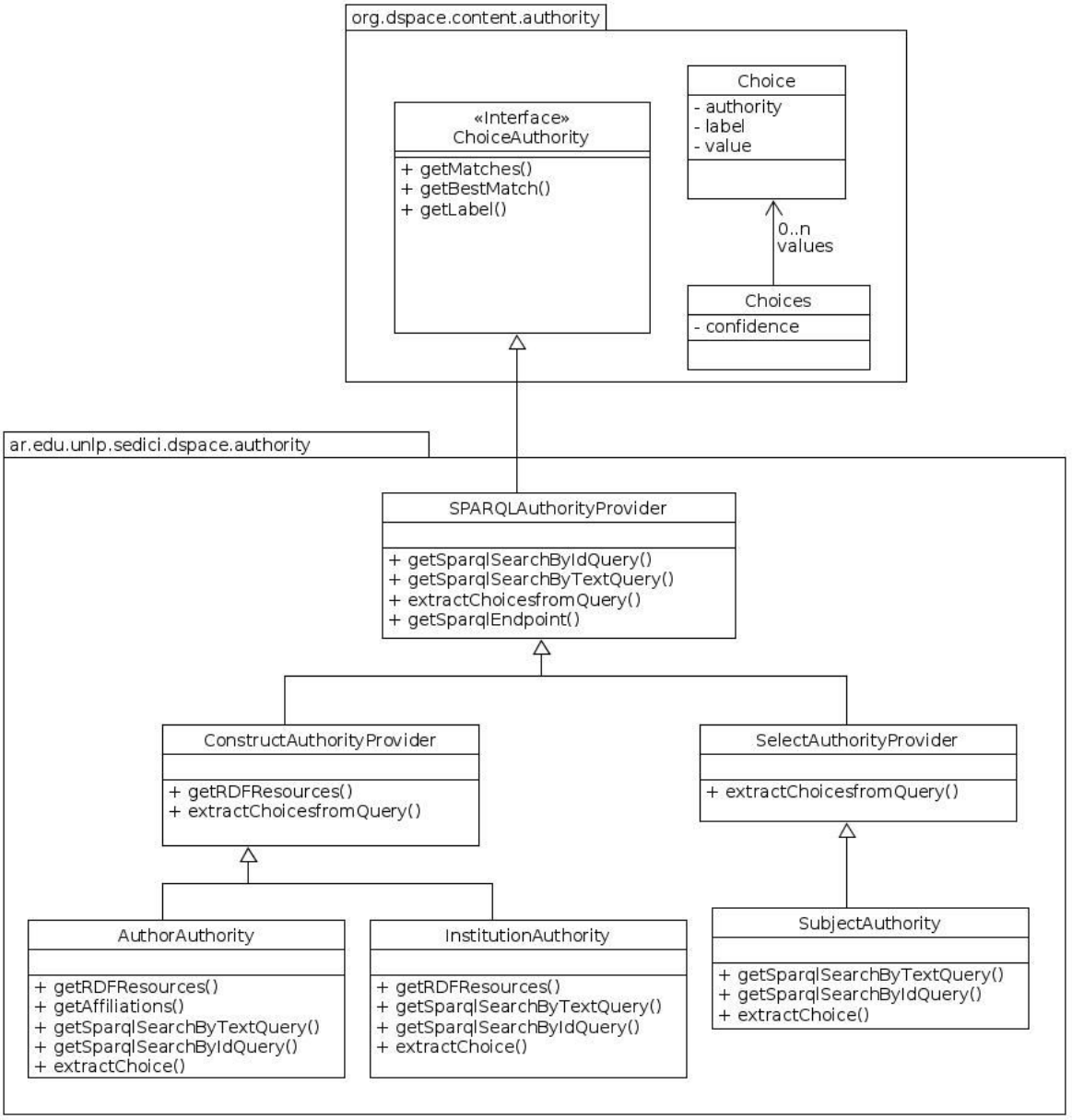

Figura 27. Diagrama de clases que extiende el módulo de autoridades de DSpace

La siguiente imagen muestra como un usuario desde DSpace realiza una consulta por el nombre del autor de una obra, mapeado con el metadato *dcterms.creator*. Como este metadato se encuentra controlado por la clase AuthorAuthority, esta será la encargada de devolver al usuario un conjunto de autoridades que coincidan con los criterios solicitados. En el método *getMatches()* se genera la consulta SPARQL que será enviada al endpoint en Drupal.

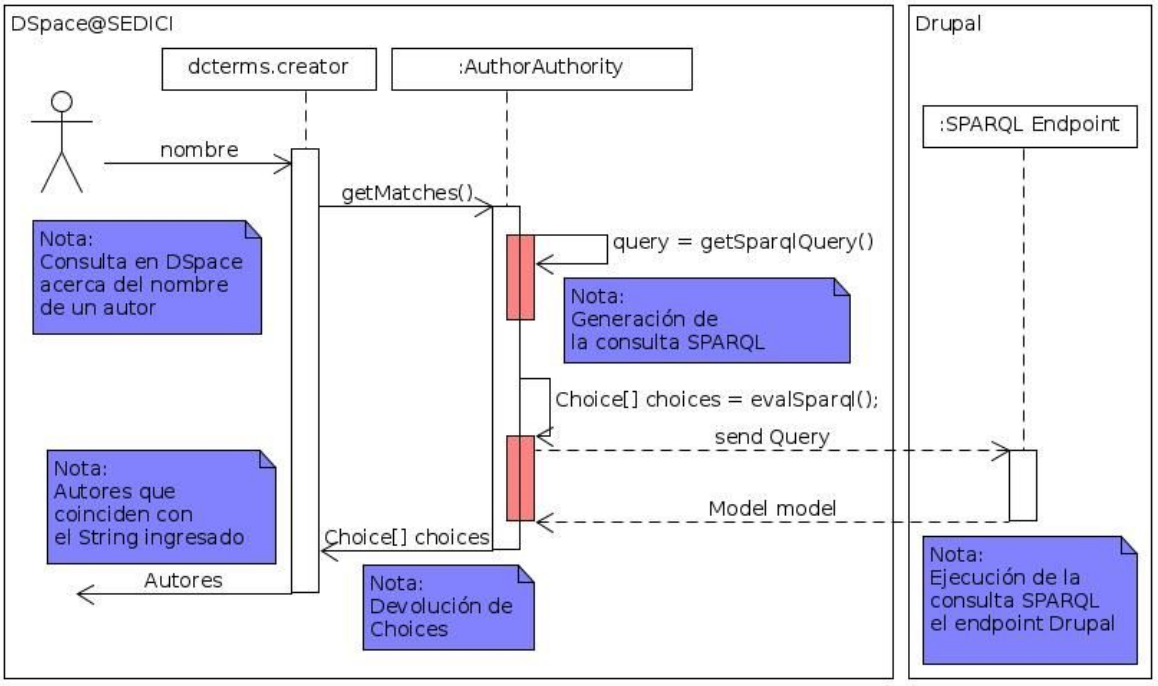

Figura 28. Diagrama de Secuencia del proceso de recuperación de un valor de autoridad

En la figura 28 se ve como un usuario hace una consulta en DSpace por el nombre de un autor. Un caso donde ocurre esto es mientras se completa el campo destinado a ingresar el nombre del autor, relacionado con el campo dcterms.creator, durante la carga de un ítem. Una vez ingresado un conjunto de caracteres, el módulo de ChoiceManagment de DSpace pide a la clase encargada de controlar el campo dcterms.creator, en este caso *AuthorAuthority*, los autores que contengan en su nombre los caracteres ingresados. Esta clase realiza la consulta SPARQL a la nueva fuente de autoridades, y en base a la respuesta generada se instancian tantos objetos *Choice* como autoridades recuperadas. Luego el módulo de ChoiceManagment se encarga de generar una vista ofreciendo al usuario seleccionar alguno de los autores que coinciden con los caracteres ingresados. Si el usuarios continúa ingresando caracteres, se volverá a realizar búsqueda en base a estos últimos. En dspace/modules/additions/src/main/java/ar/edu/unlp/sedici/dspace/authority se

encuentra disponible el código de las clases usadas para implementar los nuevos conectores de SEDICI.

Capítulo 8 - Conclusiones y trabajo a futuro

Conclusiones

A raíz del trabajo realizado se encontró en el CMS Drupal la posibilidad de desarrollar un sistema personalizado y flexible, que permitió desarrollar un módulo que gestione y exponga, a través de un endpoint SPARQL, un conjunto vocabularios controlados. Si bien existen sistemas dedicados al manejo de vocabularios controlados, como VocBench y TemaTres, estos resultaron muy especializados para las tareas que el repositorio SEDICI requería.

Para el desarrollo del sistema fue necesario comprender cómo funciona Drupal, su arquitectura y cómo es posible agregar nuevas funcionalidades a través de módulos desarrollados por la comunidad y módulos desarrollados por uno mismo a partir de los existentes. También fue necesario conocer herramientas dedicadas al desarrollo en Drupal así como buenas prácticas para gestionar el versionado del código de los módulos implementado.

Para ~~el~~ realizar el proceso de migración no sólo fue necesario pasar la información de un sistema a otro, sino que al existir datos duplicados e información no deseada se debió realizar un análisis de requerimientos que permitió distinguir qué datos debían ser normalizados o eliminados para una posterior migración.

El hecho de estructurar los vocabularios controlados en tripletas RDF para luego exponerlas través de un endpoint SPARQL implicó un acercamiento hacia el enfoque propuesto por los principios de Linked Data que facilitó el desarrollo realizado en DSpace, extendiendo la API de control de autoridades para realizar consultas SPARQL al endpoint del nuevo sistema de gestión de vocabularios controlados. Este esquema brindó una solución flexible y replicable a otros repositorios institucionales que necesiten definir autoridades para personas e instituciones, u otros tipos de vocabularios controlados, como por ejemplo una taxonomía.

El hecho de utilizar Drupal como sistema base deja abierta la posibilidad de extender y personalizar aún más los módulos desarrollados acorde a las necesidades de cada institución.

Debido a que DSpace permite implementar conectores con la funcionalidad necesaria para consumir las autoridades gestionadas en este nuevo sistema desarrollado en Drupal es posible utilizar dicho sistema como fuente de autoridades en cualquier instalación de DSpace. Uno de los repositorios institucionales que puede utilizar este sistema de gestión de vocabularios controlados es CIC-DIGITAL, el repositorio institucional de la Comisión de Investigaciones Científicas (CIC), también desarrollado sobre DSpace.

Trabajo a futuro

Realizada la migración de personas e instituciones de CelsiusDL resta definir un plan de migración para exportar al nuevo sistema de gestión de vocabularios controlados los tesauros y la jerarquía de materias y grados alcanzados. Para el caso de los tesauros habría que analizar si, una vez realizada la migración, se debe actualizar los mismos o hacer referencias a endpoints ya existentes.

En cuanto al desarrollo hecho en Drupal se deja como trabajo a futuro la integración de los módulos desarrollados en una distribución de Drupal que defina la estructura de un vocabulario controlado y permita ser extendido por un tipo de contenido existente. La implementación de una distribución en Drupal permitiría proponer su uso a la comunidad en otros repositorios y fomentar su uso. Acorde a los visto anteriormente queda como trabajo a futuro replicar en CIC-DIGITAL el sistema desarrollado para SEDICI para poder definir autores e instituciones

Otro de los puntos que se pueden trabajar es el modelado y la gestión de variantes para las autoridades en AuthVoc, de esta forma una persona o institución pueda ser referenciada a través de varios nombres, por ejemplo para el autor “Juan González” se puede definir la variante “J. Gonzalez”.

Queda abierta la posibilidad de implementar en DSpace nuevos conectores que consuman información de otros sistemas a través de SPARQL, por ejemplo el tesauro de la UNESCO que tiene disponible al público un endpoint de consulta. Al haber mapeado los vocabularios controlados al formato RDF y haber utilizado SPARQL como lenguaje de intercambio de información queda pendiente la posibilidad de mejorar la interoperabilidad semántica entre el repositorio y otros sistemas dando soporte a los mismos para que puedan, en un futuro, exponer los recursos almacenados, como artículos, tesis, revistas de acuerdo a las premisas de Linked Data.

Bibliografía

«About AGROVOC | Agricultural Information Management Standards (AIMS)». Accedido 11 de junio de 2018. http://aims.fao.org/standards/agrovoc/concept-scheme.

«About Taxonomy». Drupal.org, 18 de abril de 2010. https://www.drupal.org/docs/7/organizing-content-with-taxonomies/about-taxonomies.

«AGROVOC Linked Open Data | Agricultural Information Management Standards (AIMS)». Accedido 28 de junio de 2018. http://aims.fao.org/standards/agrovoc/linked-data.

«Architecture - DSpace 6.x Documentation - DuraSpace Wiki». Accedido 17 de junio de 2018. https://wiki.duraspace.org/display/DSDOC6x/Architecture.

«Art & Architecture Thesaurus (Getty Research Institute)». Accedido 11 de junio de 2018. http://www.getty.edu/research/tools/vocabularies/aat/.

BARTOC. «About | BARTOC.org». Accedido 11 de junio de 2018. https://bartoc.org/en/content/about.

Bernal, Isabel. «Uso de Vocabularios Controlados en Repositorios. La experiencia de DIGITAL.CSIC», 24 de noviembre de 2016. https://digital.csic.es/handle/10261/140742.

Berners-Lee, Tim. «Linked Data - Design Issues». Accedido 11 de junio de 2018. https://www.w3.org/DesignIssues/LinkedData.html.

Biblioteca do Museu de Arte de São Paulo (MASP). «Controle de Autoridades | BARTOC.org». Accedido 11 de junio de 2018. http://bartoc.org/en/node/18610.

Biblioteca Nacional de España. «5.6. Encabezamientos de materia». Accedido 11 de junio de 2018. http://www.bne.es/es/Micrositios/Publicaciones/AUTORIDADES/005_Registros/006_Encabezamientos/.

Biblioteca Universidad De Salamanca. «Catálogo “Biblioteca Universidad De Salamanca”». Accedido 11 de junio de 2018. https://bibliotecas.usal.es/catalogo-de-autoridades.

CANTIC. «CANTIC: Què és». Accedido 11 de junio de 2018. http://cantic.bnc.cat/quees.

Carrrascosa. «Tesauros y ontologias». Accedido 13 de junio de 2018. http://personales.upv.es/ccarrasc/doc/2003-2004/tesaurosonto/principal.html.

COAR. «FAQs for Controlled Vocabularies». Accedido 11 de junio de 2018. https://www.coar-repositories.org/activities/repository-interoperability/coar-vocabularies/controlled-vocabularies-faq/.

Congress, Library of, y PREMIS Editorial Committee. «PREMIS OWL Ontology (PREMIS, Preservation Metadata Maintenance Activity, Library of Congress)». Webpage. Accedido 11 de junio de 2018. http://www.loc.gov/standards/premis/ontology/index.html.

DBLP. «dblp: How to use the dblp search API?» Accedido 11 de junio de 2018. http://dblp.org/faq/How+to+use+the+dblp+search+API.html.

———. «dblp: What is dblp?» Accedido 11 de junio de 2018. http://dblp.org/faq/What+is+dblp.html.

«drupal_flow_0.gif (528×562)». Accedido 29 de junio de 2018. https://www.drupal.org/files/drupal_flow_0.gif.

DuraSpace. «Authority Control of Metadata Values - DSpace - DuraSpace Wiki». Accedido 17 de junio de 2018. https://wiki.duraspace.org/display/DSPACE/Authority+Control+of+Metadata+Values.

«EuroVoc». Accedido 13 de junio de 2018. http://eurovoc.europa.eu/drupal/?q=es/node.

«FOAF Vocabulary Specification». Accedido 11 de junio de 2018. http://xmlns.com/foaf/spec/.

«Getty Vocabularies». Accedido 11 de junio de 2018. http://vocab.getty.edu/.

Lapuente, María Jesús Lamarca, y Chusa Lamarca Lapuente. «Ontologías». Tesis. Accedido 13 de junio de 2018. http://www.hipertexto.info/documentos/ontologias.htm.

«Learn RDF». *Cambridge Semantics* (blog). Accedido 11 de junio de 2018. https://www.cambridgesemantics.com/blog/semantic-university/learn-rdf/.

MeSH. «Medical Subject Headings - Home Page». Product, Program, and Project Descriptions. Accedido 11 de junio de 2018. https://www.nlm.nih.gov/mesh/meshhome.html.

———. «MeSH Linked Data». Accedido 11 de junio de 2018. https://id.nlm.nih.gov/mesh/.

Narvaez, Efren, y Nelson Piedra. «Un enfoque de Linked Data para garantizar la interoperabilidad semantica e integridad de datos academicos universitarios», s. f., 13.

OCLC. «VIAF». Accedido 11 de junio de 2018. https://www.oclc.org/es/viaf.html.

«Organizing Content with Taxonomy». Drupal.org, 7 de junio de 2002. https://www.drupal.org/docs/7/organizing-content-with-taxonomies/organizing-content-with-taxonomy.

«Owl 101 - Cambridge Semantics | Cambridge Semantics». Accedido 11 de junio de 2018. https://www.cambridgesemantics.com/blog/semantic-university/learn-owl-rdfs/owl-1

01/.
«Portal SNRD». Accedido 11 de junio de 2018. http://repositoriosdigitales.mincyt.gob.ar/dnet-web-generic/.
«RDF - Semantic Web Standards». Accedido 11 de junio de 2018. https://www.w3.org/RDF/.
«SIOC Core Ontology Specification». Accedido 11 de junio de 2018. http://rdfs.org/sioc/spec/.
«SKOS». Accedido 11 de junio de 2018. http://skos.um.es/acerca/index.php.
«STW Thesaurus for Economics: Home». Accedido 11 de junio de 2018. http://zbw.eu/stw/versions/latest/about.
«TemaTres: servidor de vocabularios controlados». Accedido 15 de junio de 2018. http://r020.com.ar/tematres/manual/.
Texidor, Silvia. «Control de autoridades». Accedido 11 de junio de 2018. http://www.bnm.me.gov.ar/redes_federales/bera/encuentros/nacionales/2008_ref/docs/texidor.pdf.
The Library of Congress. «LC Linked Data Service: Authorities and Vocabularies (Library of Congress)». Webpage. Accedido 11 de junio de 2018. http://id.loc.gov/authorities/subjects.html.
«UDC Consortium - About UDC». Accedido 28 de junio de 2018. http://www.udcc.org/index.php/site/page?view=about.
«UDC Summary Linked Data». Accedido 28 de junio de 2018. http://udcdata.info/.
«Vista General del Lenguaje de Ontologías Web (OWL)». Accedido 11 de junio de 2018. https://www.w3.org/2007/09/OWL-Overview-es.html.
«VocBench: A Collaborative Management System for SKOS-XL Thesauri». Accedido 15 de junio de 2018. http://vocbench.uniroma2.it/.
Wolf, Gunnar. «Interoperabilidad: ¿A qué aspiramos cuando hablamos de ella? | SG Buzz». Accedido 13 de junio de 2018. https://sg.com.mx/revista/33/programar-es-un-estilo-vida-interoperabilidad.
«YASGUI». Accedido 28 de junio de 2018. http://about.yasgui.org/.