DCOntoRep: hacia la interoperabilidad semántica de Repositorios Institucionales de Acceso abierto

Sandobal Verón, Valeria C.; Ale, Mariel; Gutiérrez, María de los Milagros

GIESIN, UTN FRR Resistencia, H3500CHJ, Argentina – vsandobal@frre.utn.edu.ar
CIDISI, UTN FRSF, Santa Fe, S3004EWB, Argentina

Resumen. Debido a la heterogeneidad de los sistemas que sirven de repositorio para el almacenamiento de objetos digitales, ha surgido con interés el desarrollo de ontologías que permita interpretar de manera correcta el significado de los conceptos involucrados. Los objetos digitales depositados en repositorios deben ser etiquetados, catalogados y clasificados y para ello se utilizan los llamados metadatos, que proveen información sobre los objetos digitales. Es así que surgen estándares de metadatos tales como Dublin Core (DC) para la descripción de recursos web, Learning Object Model (LOM) para la descripción de recursos educativos específicamente; entre otros. Si se considera que se busca la comunicación entre los repositorios existentes, se hace necesaria la correcta interpretación semántica de los conceptos involucrados en los metadatos. La adopción de uno de estos estándares para la implementación de un repositorio depende de los objetivos que la institución ejecutora persiga. El presente trabajo propone una ontología para el estándar DC con las directrices del Sistema Nacional de Repositorios Digitales (SNRD). Teniendo en cuenta que el estándar DC es el más utilizado por los repositorios actualmente vigentes y las directrices SNRD lo utilizan a fin de poder adherir un futuro repositorio institucional a los cosechados por el Ministerio de Ciencia, Tecnología e Innovación Productiva

Keywords: objetos digitales, estándares de metadatos, Repositorios institucionales de acceso abierto, Dublin Core, recomendaciones SNRD

1 Introducción

En los últimos años se ha visto un incremento considerable en el uso e implementación de Repositorios institucionales de acceso abierto a partir de la promulgación de la ley 26899[1] que exige a los organismos e instituciones públicas que forman parte del Sistema Nacional de Ciencia y Tecnología publicar en acceso abierto, su producción científica – tecnológica. En este contexto, surge un extenso conjunto de problemáticas que deben ser resueltas, entre las cuales se puede

[1] Sitio Web Ministerio de Ciencia, Tecnología e Innovación Productiva. Ley 26899.http://repositorios.mincyt.gob.ar/recursos.php

mencionar las tecnológicas, relacionadas con la implementación, funcionamiento y uso. Este trabajo hace su aporte en el área de la interoperabilidad, específicamente en la interoperabilidad semántica de la información almacenada.

El valor de un repositorio institucional de acceso abierto es su potencial para interoperar con otros repositorios formando una red de repositorios la cual configura una infraestructura de e-investigación [1]. Para lograr este objetivo, surge la necesidad de estandarizar la manera de describir los objetos almacenados, de forma tal que permitan búsquedas más concretas de los recursos disponibles, como así también la reutilización de los mismos. Es así que surgen estándares de metadatos tales como el Dublin Core (DC – Metadata Element Set)[2] que permite la descripción de cualquier tipo de recurso disponible en la web y LOM Learning Object Model [3] el cual se utiliza para la descripción de recursos educativos específicamente. Sin embargo el simple uso de estos estándares no garantiza la correcta localización y descripción de los recursos, por lo tanto el Sistema Nacional de Repositorios Digitales (SNRD) ha establecido una normativa donde se especifica la forma en que dichos metadatos deben ser completados [4].

El objetivo de este trabajo es aportar una herramienta semántica, específicamente una ontología denominada *DCOntoRep*, que dé soporte a la descripción de objetos de aprendizajes (LO por su sigla en inglés Learning Object) de una manera estandarizada, considerando no sólo el estándar DC sino también incorporando las directivas de SNRD a través de la definición de axiomas, reglas de derivación e instancias en la ontología propuesta. La misma será implementada en el Repositorio Institucional (RI) que se obtenga como resultado del proyecto de investigación que actualmente se está llevado a cabo en centro de investigación CIDISI – Facultad Regional Santa Fe: "Desarrollo e Implementación de un repositorio institucional de acceso abierto para objetos digitales educativos". Dado que dicho repositorio se plantea para albergar objetos digitales educativos, es decir aquellos generados desde la parte académica de la universidad, como así también desde el área de investigación y tecnología, es que en este trabajo se considera apropiado utilizar el concepto de LO, teniendo en cuenta que el mismo puede definirse como "una entidad, digital o no digital, que puede ser utilizada, reutilizada y referenciada durante el aprendizaje apoyado con tecnología" [5], o como lo define García Aretio [6]: "archivos o unidades digitales de información dispuestos con la intención de ser utilizados en diferentes propuestas y contextos pedagógicos."

Se define el objetivo de ***DCOntoRep*** como: "***representar la semántica explícita en las etiquetas utilizadas por el estándar DC y las directrices SNRD que permita mejorar la búsqueda, reutilización y depósito de LO, en los repositorios institucionales que utilizan este estándar***".

2 Desarrollo de DCOntoRep

Una ontología es una especificación formal y explícita de una conceptualización compartida [7]. Por lo tanto debe tener una especificación, una definición y debe ser explícita teniendo en cuenta el contexto en el que se la está utilizando; además de que

la misma es compartida, por lo cual debe ser entendida por al menos dos actores involucrados en la especificación.

La ontología que a continuación se presenta, tiene como base de información un archivo semiestructurado XML. Según [8] los métodos propuestos para el aprendizaje de una ontología se pueden clasificar en: aprendizaje desde corpus de texto, aprendizaje desde instancias, aprendizajes desde esquemas y aprendizajes desde mapeos semánticos. Siguiendo la propuesta realizada por [9] a partir de archivo XML generado con las etiquetas del estándar DC se aplican las heurísticas definidas tales como: las clases OWL surgen de tipos complejos (<xsd:complexType>); la figura 1(a) muestra esta heurística con el tipo complejo *LearningObject.* Las clases OWL surgen de elementos a nivel de esquema de un tipo complejo; la figura 1(b) muestra esta heurística con el tipo complejo *Title* que es parte del esquema DCMetadata. Los *Object Property* (relaciones) en OWL, surgen a partir de las relaciones entre elementos y subelementos del archivo XML, como se muestra en la figura 2 con la relación *hasSubject.* Los *Data Property* (atributos) de OWL surgen de tipos simples, tales como xs:date, xs:string, mostrado en la figura 3.

En este mapeo resulta necesario, algunas otras conversiones tale como: facetas para restringir valores de elementos simples como *Enumeration*, que pueden ser definido en OWL2 como valores de rango para los data property y los valores de *MinOccurs* y *MaxOccurs* se especifican como valores que deben ser cumplidos en la estructura del *Equivalent to*

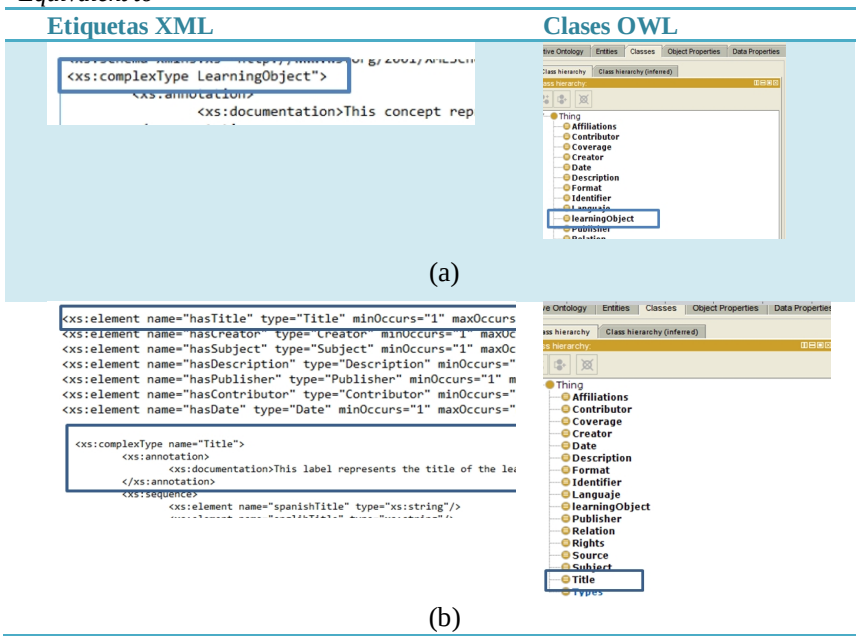


Fig. 1. Mapeos entre elementos de XML a clases OWL

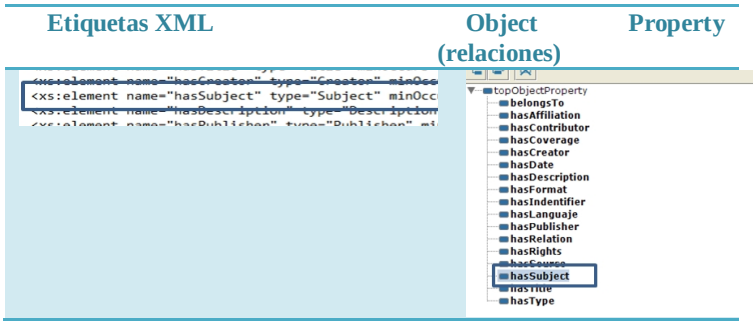


Fig. 2. Etiquetas XML mapeadas a relaciones OWL

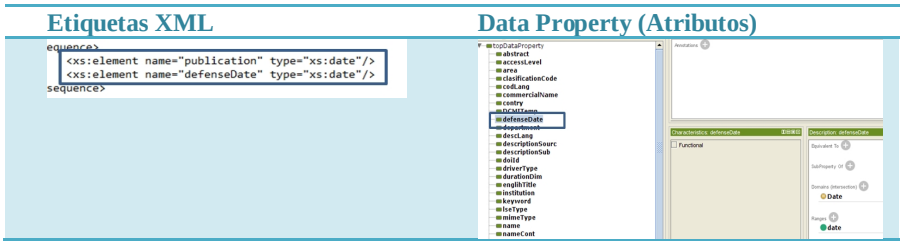


Fig. 3. Etiquetas XML a Data Property OWL

Luego del mapeo propuesto desde el schema XML se realizaron los pasos sugeridos por el método basado en Meta-Modelo de Ingeniería de Procesos de Software y Sistemas, versión 2.0, Software & System Process Engineering Meta-Model, SPEM.0 [10]. En su desarrollo, se tuvieron en cuenta las directrices SNRD y se formularon algunas preguntas de competencia que permitan la validación. Posteriormente, se analizaron las directivas del SNRD para enriquecer la ontología agregando relaciones, axiomas y reglas que fueran necesarias.

El gráfico de la figura 4 muestra los principales conceptos definidos en la ontología *DCOntoRep.* Se define el concepto *LearningObject* para representar los objetos de aprendizaje que son descriptos a través de metadatos. El concepto *DcMetadata* abstrae los metadatos de DC. Como subclases de este concepto se definieron tres clases correspondientes a cada una de las categorías especificadas en DC: (i) *Content*, que agrupa los metadatos para describir el contenido de un recurso, (ii) *IntellectualProperty* que agrupa los metadatos que describen las cuestiones referentes a la propiedad intelectual y derechos de uso del recurso y finalmente (iii) *Instantiation* correspondiente a los metadatos que describen cuestiones técnicas de la instancia del recurso que describen.

Como subclases de *content* se definieron los conceptos: *Title*, para contener el título del LO; *Subject*, que representa el dominio del LO; *Description* permite describir el LO a través de un texto significativo; *Source*, identifica un objeto digital que se ha tenido en cuenta para la formulación del LO; *Language* identifica el lenguaje

correspondiente al LO*; Relation* permite identificar otras versiones del LO que se describe; *Coverage* se refiere al alcance o ámbito del LO, generalmente incluye la ubicación espacial, un período de tiempo o la jurisdicción. Para el concepto *IntellectualProperty* se representaron las siguientes subclases: *Contributor* que identifica el/los colaborador/es que hayan aportado al contenido del LO, el cual puede ser una entidad o una persona con cargo como por ejemplo directores, supervisores, editores, entre otros; *Creator* que identifica el/los autor/es principal del LO; también puede designar una institución o un evento; *Publisher* corresponde al publicador o editor que hace posible que el recurso esté disponible (puede ser una persona, una organización o un servicio); *Right* identifica los derechos de autor asociados al LO. Luego como subclases de *Instantiation* se identificaron los siguientes conceptos: *Date* que especifica la fecha de creación o disponibilidad del LO*; Type* que establece el tipo de resultado científico, como puede ser un artículo o un libro entre otros*; Format* que identifica el formato físico o digital del LO*;* e *Identifier* hace referencia a una URL o URI que identifica unívocamente al LO.

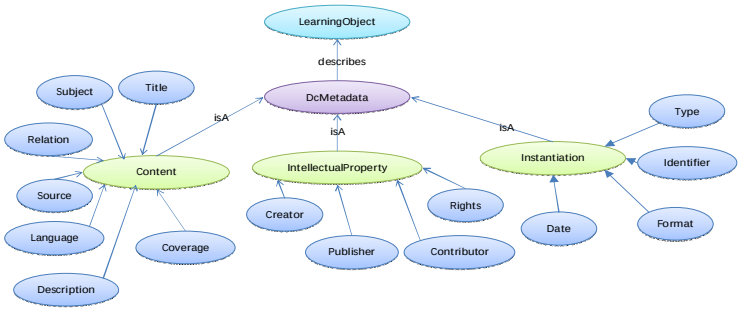
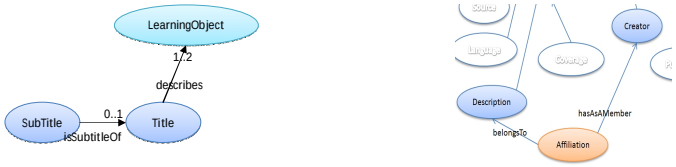


Fig. 4. Ontología DCOntoRep

2.1 Enriquecimiento de la ontología

El principal objetivo del enriquecimiento de la ontología es el mejoramiento en la representación de las entidades de la ontología base modeladas según el dominio. Para realizar esta actividad se tomaron las directrices impuestas por el SNRD. De esta forma, para cada uno de los metadatos, se establece cuestiones relacionadas a la obligatoriedad o no de dicho metadato, la posibilidad de repeticiones y el formato del contenido de los metadatos. Así por ejemplo se determina que el metadato *Title* es obligatorio, en caso de que exista subtítulo debe estar separado del título por dos puntos y si la obra tiene títulos en distintos idiomas deben aparecer dos instancias del metadato, una para cada título. Para reflejar esta restricción se agregó a la ontología el concepto *SubTitle* relacionado con *Title* a través de la relación *isSubTitleOf* (relación inversa *hasSubTitle)* como se muestra en la figura 5(a). Las restricciones de repetición y obligatoriedad se implementaron a través de restricciones de cardinalidad.



(a) (b)

Fig. 5. Modificaciones a la ontología según directrices SNRD

Siguiendo las directrices del SNRD, la etiqueta *description* debe especificarse también las afiliaciones de cada uno de los autores del LO. La afiliación corresponde a la institución/universidad a la que pertenecen los autores. Para dar cumplimiento a estas indicaciones, tal como se muestra en la figura 5(b), se agregó la clase *Affiliation*, la cual se encuentra relacionada con la clase *Creator* a través de la relación *hasAsAMember* y con la clase *Description* a través de la relación *belongsTo*. Affiliation tiene dos atributos: *institution* y *department*. Para representar la restricción de que todo autor tenga una afiliación correspondiente, se definieron axiomas de integridad como muestra la expresión lógica (1).

$$CreatorWithAffiliation \equiv Creator \cap \exists hasAffiliation.Affiliation \cap \forall hasAffiliation.Affiliation \quad (1)$$

Este mismo proceso de enriquecimiento se siguió para todas las directrices definidas por el SNRD. Las mismas recomiendan el uso de estándares para completar metadatos tales como la norma ISO 639[2](metadadato *language)* y la norma ISO 3166[3](metadato *coverage)*. Para reflejar esto, se importaron dos ontologías que conceptualizan dichas normas.

En el caso de la ontología importada ISO639 se encuentra el concepto *Language* con dos atributos *alpha2*(ISO 639-1) y *alpha3*(ISO 639-2). Las directrices SNRD recomienda el uso de la ISO 639-3 para la codificación de la etiqueta *Language* representada por el atributo *alpha3* en la ontología ISO639.

Luego, se eliminó la clase *Language* de la ontología DCOntoRep y se la sustituye con la clase *Language* de la ontología importada ISO639, dejando como atributo solo *alpha3*. En la Figura 6 se muestra el antes y después de este proceso de enriquecimiento.

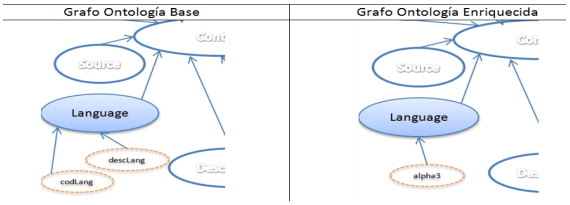


Fig. 6. Comparación Ontología Base – Ontología Enriquecida. Clase *Language*

[2] http://www.iso.org/iso/es/home/standards/language_codes.htm

[3] http://www.iso.org/iso/home/standards/country_codes.htm

Para el caso de la ontología ISO3166 al importarla se incorpora la clase *Country* y los atributos *code* y *name*; además para relacionar la clase *Coverage* con la nueva clase *Country* se agrega la relación *coversA* (Figura 7). Según las directrices del SNRD la ISO 3166 a utilizar es la 2, que define los códigos de identificación de las principales subdivisiones (provincias o estados) de todos los países codificados en ISO 3166-1; así Argentina, Provincia del Chaco, estaría codificado como AR-H.
Para la etiqueta *type* es necesario agregar los atributos que responden a los vocabularios controlados sugeridos por el SNRD; se agrega: *driver*: resultado científico en términos del vocabulario controlado DRIVER; *snrdType*: subtipo del resultado científico acordados por el SNRD; *version*: indica la versión del LO según el vocabulario controlado DRIVER(*draft, submittedVersion, acceptedVersion, publishedVersion* y *updateVersion)*. Las instancias de *driver* y *snrdtype* se establecen como obligatorias y la de *version* como opcional, todos estos casos se reflejan a través de restricciones de cardinalidad, tal como se muestra en la figura 8

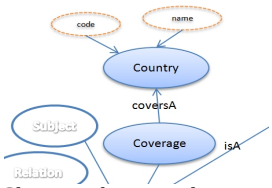


Fig. 7. Clases, atributos y relaciones agregadas al importar la ontología ISO 3166-2

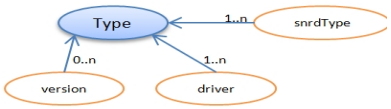


Fig. 8. Restricciones de Cardinalidad para los atributos de la clase Type

A la ontología enriquecida, se le ha sumado reglas con el uso de SWRL que permiten clarificar ciertas reglas de negocio que no han podido ser expresadas a través de clases, atributos, relaciones, y que deben formar parte del conocimiento de la ontología.
Según las directrices SNRD, si un LO es del tipo *doctoralThesis*, *masterThesis* o *bachelorThesis* para el atributo *driver*; es obligatorio incluir al menos un colaborador, definido como director de la tesis, lo cual se especifica en la regla(2):

$$LO(?l) \wedge (hasType(?t,\text{"doctoralThesis"}) \vee hasType(?t,\text{"masterThesis"}) \vee hasType(?t,\text{"bachelorThesis"})) \rightarrow hasContributor(?l, ?t) \quad (2)$$

Las directrices SNRD establece que para cada tipo de la clasificación *driver* corresponde un subconjunto de tipos establecidos por el propio SNRD, por ejemplo si un LO es *bachelorThesis* en *driver*, corresponderá al tipo *tesis de grado* o *trabajo final de grado* para el tipo snrd, tal como se muestra en la expresión (3).

$$Type(?t) \wedge driver(?t, \text{"bachelorThesis"}) \rightarrow snrd(?t,\text{"tesis de grado"}) \vee snrd(?t,\text{"trabajo final de grado"}) \quad (3)$$

Si *driver* es *review*, corresponderá al tipo *reseña artículo* o *revisión literaria* para el tipo snrd como se muestra en la expresión (4).

$$Type(?t) \wedge driver(?t, \text{"review"}) \rightarrow snrd(?t,\text{"reseña artículo"}) \vee snrd(?t,\text{"revisión literaria"}) \quad (4)$$

Así mismo, podríamos agregar reglas que cierren el mundo sobre el cual se está trabajando, donde por ejemplo se establezca que si un LO tiene dos autores y esos auto-

res son instancias diferentes entre sí, uno de ellos es colaborador del autor del LO, tal como se define en la expresión (5).

LO(?l) ^ hasCreator(?l, ?c1) ^ hasCreator(?l, ?c2) ^differentFrom (?c1, ?c2)
-> hasContributor(?c1, ?c2) (5)

3 Caso de estudio.

Con el objetivo de validar la ontología obtenida se pobló la misma con instancias que permitan ejecutar consultas en SPARQL que respondan a las preguntas de competencia presentadas en la Tabla 1. Para ello se tomó como ejemplo un documento de conferencia cuyo autor es *Sandobal* perteneciente al grupo de investigación *GIESIN* de la *UTN-FRRe*. Como colaborador se encuentra *Gutiérrez*, que cumple el rol de director. La publicación del artículo está a cargo de la *UTN-FRRe*, en particular de la Secretaría de Ciencia y Tecnología. Los derechos que se han establecidos en relación al nivel de acceso es de acceso libre (open Access, según la definición establecida por el SNRD) y la uri en la que se basa este tipo de licencia es http://creativecommons.org/licenses/by/2.5/ar/, las instancias que se muestran en la figura 9 correspondientes a la categoría *Intellectual Property*. En la figura 10 se puede visualizar las instancias de la categoría *Content*, donde el título del artículo es "*Hacia la integración*". El idioma es *Español*, por lo cual el código alpha3 es *spa*. Una de las fuentes que se ha tenido en cuenta para el desarrollo del artículo se ha identificado con el ISBN: 978-950-42-0142-7. Un documento relacionado estás identificado con la URI http://www.ssoar.info/es/home/sobre-ssoar.htm. El dominio está definido por las palabras claves repositorio y acceso abierto. El ámbito está determinado como en *Argentina-Chaco*, donde el código es *AR-H*. Para la categoría *Instantation*, que se muestra en la figura 11, se estableció como fecha de publicación *2014-06-12*, el formato según la clasificación *Mime Type* corresponde a *text*, el tipo es *documento de conferencia*, según Driver corresponde a *document conference* y la versión es *accepted*. La URL a la cual se puede ingresar para tener acceso al artículo es http://www.frre.utn.edu.ar/secyt/paginas/view/item/ii_jornadas_de_investigacion_en_ingenieria_del_nea_y_paises_limitrofes.

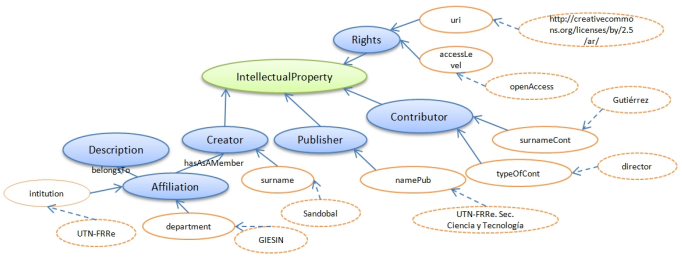


Fig. 9. DCOntoRep – Subclase *IntellectualProperty* y sus instancia

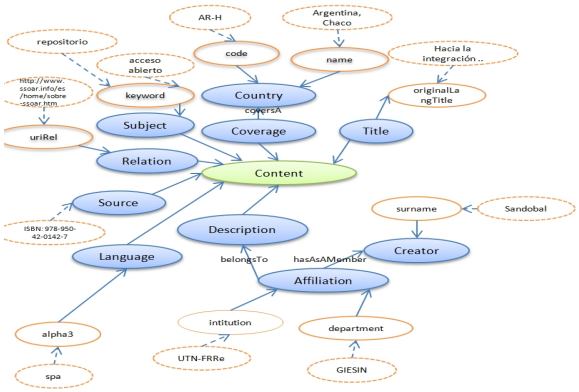


Fig. 10. DCOntoRep – Subclase *Content* y sus instancias

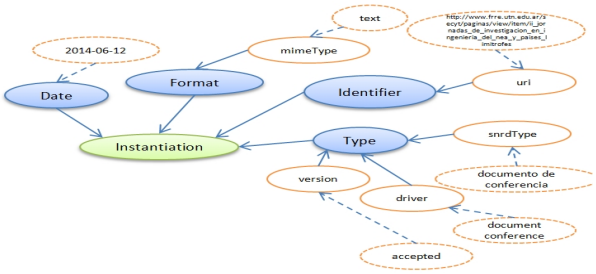


Fig. 11. DCOntoRep – Subclase *Instantiation* y sus instancias

En la tabla 2 se muestran las preguntas de competencias con sus correspondientes consultas SPARQL

Preguntas de Competencia	Consulta SPARQL
¿Cuáles son los objetos de aprendizaje cuyo autor es Y?	SELECT ?originalLangTitle WHERE {dc:Sandobal dc:isCreatorOf ?originalLangTitle} Hacia Evaluacion
¿Cuál es el tipo según SNRD del objeto de aprendizaje X cuyo autor es Y?	SELECT ?snrd WHERE {dc:Sandobal dc:isCreatorOf dc:Evaluacion . dc:Evaluacion dc:hasTypeOfSnrd ?snrd} DocumentoConferencia

Tabla 2. Preguntas de Competencias y sus correspondientes consultas SPARQL

4 Conclusiones

El presente artículo presenta una ontología basada en el estándar DC considerando las directrices SNRD para la utilización en la descripción de LO que se encuentran en repositorios institucionales. El trabajo consistió en entender el significado de cada una de las etiquetas propuestas por el estándar DC y su adecuación en las directrices SNRD para proponer una ontología que responda a la misma. Al contar con la descripción específica de los metadatos, se recurrió a las preguntas de competencia para que sirvan de guía en la validación de la ontología obtenida. Así mismo, en el documento de especificación de requerimientos se planteó que la ontología dé soporte a: (i) la búsqueda de objetos digitales descriptos a través de metadatos en DC, (ii) guiar al usuario en la correcta introducción de los valores de las etiquetas de los objetos digitales.

A partir de las consideraciones antes mencionadas y las evaluaciones realizadas de las preguntas de competencias ejecutadas en SPARQL es posible afirmar que la ontología responde a los requerimientos especificados. La ontología obtenida se considera una aproximación hacia la construcción de ontologías para el uso de metadatos, que ayude a definir un mapeo ontológico entre estándares de metadatos. Este trabajo forma parte de un proyecto que tiene como objetivo la implementación de un repositorio institucional en UTN – FRSF que se encuentra en la fase de prototipo. Como trabajo futuro, se utilizará ésta ontología en el prototipo para su validación y verificación.

Referencias

[1]COAR: Confederation of Open Access Repositories. The Current State of Open Access Repository Interoperability. Working Group 2: Repository Interoperability. Octubre 2012.

[2]National information standards organization, “The Dublin core metadata element set”, ISSN 1041-5635, 2013

[3] IEEE Standard for Learning Object Metadata. (2002). Learning Technology Standards Comittee of the IEEE Computer Society. 1484.12.1-2002

[4] SNRD (2013). Directrices SNRD: Directrices para proveedores de contenido del Sistema Nacional de Repositorios Digitales Ministerio de Ciencia, Tecnología e Innovación Productiva.

[5] IEEE. Learning Technology Standards Committee, 2002, p 45.

[6] García Aretio. “Objetos de Aprendizaje”. Bol. Elect. de Noticias de Educación a Distancia – Bened. Disponible en: http://e-spacio.uned.es/fez/eserv.php?pid=bibliuned:329&dsID=editorialfebrero2005.pdf

[7] Studer R., Benjamins VR, Fensel D. Knowledge Engineering: Principles and Methods. IEEE Transaction and Data and Knowledge Engineering 25(1-2) :161-197. 1998

[8] Giraldo, Gloria; Marín Juan C.; Urrego Giraldo, Germán. Extracción de elementos de una ontología de dominio a partir de documentos esquema. Revistas Avances en Sistemas e Informática. Vol 6 Nº2, Septiembre 2009. Medellín. ISSN1657-7663

[9] Yahia, Nora; Moktar, Sahar; Wahab Ahmed, Abdel. Automatic Generation of OWL Ontology from XML Data Source.(2012). CoRR, abs/1026.0570

[10] Software & System Process Engineering Metamodel Specification (SPEM). Version 2.0 Object Managment Group, 2008. www.omg.org/spec/SPEM/2.0/