Una aproximación basada en Linked Data para la detección de potenciales redes de colaboración científica a partir de la anotación semántica de producción científica: Piloto aplicado con producción científica de investigadores ecuatorianos

Nelson Piedra, Janneth Chicaiza, Elizabeth Cadme, Richar Guaya

Universidad Técnica Particular de Loja, San Cayetano Alto S/N, Loja, Ecuador, 1101608.

Autor para correspondencia: nopiedra@utpl.edu.ec

Fecha de recepción: 21 de septiembre de 2014 - Fecha de aceptación: 17 de octubre de 2014

RESUMEN

En este documento se propone un marco de trabajo basado en tecnologías de la Web Semántica para detectar potenciales redes de colaboración, mediante el enriquecimiento semántico de artículos científicos producidos por investigadores que publican con afiliaciones ecuatorianas. El marco de trabajo se describe a través de un ciclo de publicación de datos enlazados. Como alcance se consideraron publicaciones que tienen al menos un autor con afiliación ecuatoriana. Las redes de colaboración detectadas son un insumo importante para fortalecer los esfuerzos del gobierno ecuatoriano y las autoridades universitarias del país, priorizar los esfuerzos y recursos invertidos en investigación y determinar la pertinencia o coherencia de los programas de investigación.

Palabras clave: Datos enlazados, red de colaboración, Ecuador, publicaciones, SCOPUS, Web semántica, sistemas de organización de conocimiento.

ABSTRACT

In this paper, a framework based on semantic Web technologies is presented to detect potential collaboration networks. The collaborators are identified through a semantic enrichment of scientific articles produced by researchers who publish with Ecuadorian affiliations. The scope of this work includes the papers, which have at least one author with Ecuadorian membership. Detected collaborative networks are an important basis to promote the efforts of the Ecuadorian government and higher education institutions, in order to prioritize the efforts and the resources invested in research and determine the relevance and coherence of the research programs.

Keywords: Linked data, collaboration network, Ecuador, papers, SCOPUS, semantic Web, knowledge organization system.

1. INTRODUCCIÓN

La producción científica refleja los resultados de una actividad académica o investigativa; en este sentido, también refleja el esfuerzo y los temas en los que se involucran las personas. Por tanto, es importante que se direccionen los recursos y se visibilice el trabajo académico y científico en relación a una determinada línea, de tal manera que un investigador pueda encontrar pares o redes de colaboración con las cuales compartir el desarrollo de una determinada área.

Sin embargo, el acceso a los datos de trabajos académicos y científicos es limitado, en varios casos se requiere tener una cuenta o suscripción a revistas y bases de datos científicas para recuperar la información requerida. Por otra parte, los trabajos están dispersos en diferentes fuentes o repositorios que no se comunican, por tanto, resulta complicado poder consultar, integrar y comparar información.

En este trabajo, nos centramos en este segundo escenario, y proponemos el uso de tecnologías semánticas y datos enlazados para mejorar la interoperabilidad de fuentes de datos y para detectar comunidades científicas.

Y es que en el escenario académico y tecnológico actual se está apostando por la publicación de datos abiertos mediante tecnologías de la Web Semántica y la aplicación de guías de Linked Data; se espera mejorar el acceso, la integración, el reusó, y la compartición de datos entre repositorios heterogéneos. Además, mediante tecnologías de Web Semántica, las máquinas pueden integrar, comparar y analizar datos codificados con estas tecnologías, y las personas pueden beneficiarse de una nueva gama de aplicaciones que les ayude a tomar mejores decisiones.

Por tanto, el objetivo de publicar la producción científica del país como datos enlazados abiertos es mejorar la visibilidad y el acceso a datos como: áreas de investigación y tópicos alrededor de los cuales trabajan los investigadores ecuatorianos y redes de trabajo subyacentes. Más adelante, estos datos podrán ser analizados para identificar: entidades de financiamiento, factores que puedan mejorar o impulsar la investigación entre las instituciones y organizaciones del país, priorizar los esfuerzos y recursos invertidos en investigación y determinar la pertinencia o coherencia de los programas de investigación respecto de los problemas que aquejan a nuestra sociedad y entorno.

En este trabajo se presenta una alternativa para la creación de redes de colaboración a través del uso de datos enlazados, tomando como base la utilización de palabras clave (*keywords*) de artículos científicos para determinar grupos de publicaciones relacionadas cuyos autores no han trabajado en coautoría. Esto permitirá descubrir a pares académicos que trabajan en temas de investigación similares y con quienes se podría llegar a establecer redes de colaboración.

En la Sección 2 de este artículo, se describen algunos trabajos en relación a datos abiertos en el Ecuador y algunas iniciativas en el contexto de la producción científica; además las tecnologías utilizadas en la presente propuesta son introducidas. Continuando con el artículo, en la Sección 3, se explica el marco de trabajo para encontrar redes de colaboración, el cual intenta aprovechar la estructura de grafo y la posibilidad de enriquecer los recursos con semántica explícita a través de datos de fuentes de conocimiento social. Un caso que pone a prueba la propuesta es presentado en la Sección 4. Finalmente, en la Sección 5, se presentan las conclusiones del trabajo y los trabajos futuros.

2. DATOS ENLAZADOS EN EL CONTEXTO DE LA PRODUCCIÓN CIENTÍFICA

2.1. Open data y la producción científica en el Ecuador

En el Ecuador, una de las comunidades que coordina los esfuerzos e iniciativas en relación a datos abiertos, es el *Open Knowledge Foundation Ecuador* (OKF Ecuador[1]), una de las oficinas que la OKF mantiene en Sudamérica, con el objetivo de abrir el conocimiento y los datos alrededor del mundo.

En el país, existe el portal de datos abiertos de Ecuador, en el sitio de este proyecto[2], que se encuentra en su versión Beta y es promovido por OKF Ecuador, se pueden descargar algunos *datasets*. Uno de los propósitos bajo los cuales se concibió este trabajo fue facilitar el acceso de los ciudadanos a datos públicos en formatos y licencias que permitan su reutilización y la creación de aplicaciones y servicios. Hasta el momento existen 27 *datasets*: 25 relacionados a mapas, 1 de presupuestos y 1 de salud.

En cuanto a producción científica, en agosto del 2013, la Secretaría Nacional de Educación Superior, Ciencia y Tecnología del Ecuador (SENESCyT[3]) organizó el seminario internacional "Herramientas para la difusión del conocimiento científico". En este evento, algunos importantes hallazgos fueron identificados por Burque-Gámez (2013):

[1] http://ec.okfn.org/
[2] www.datosabiertos.ec
[3] http://www.senescyt.gob.ec/

1. Entre el 2003 al 2012, Ecuador registra 3649 artículos Scopus y 3573 en Web of Science. Del total de artículos indexados, 8,8% corresponden a artículos de Ciencias Sociales y Humanidades y el restante 91,2% a artículos en Ciencia y Tecnología.
2. Las áreas de investigación en las que más producción científica se registra es en Física con el 12,62% y Ecología de Ciencias Ambientales (10,58%). Del área técnica, se destacan: Ingeniería con el 4,68% y Ciencias de la Computación con apenas el 2,74%
3. A nivel Latinoamericano, los países que más destacan son: Chile, Colombia y Venezuela. Ecuador ocupa el 9no. lugar de entre 17 países de la región, es decir, "Ecuador tiene una producción científica baja en el contexto latino e iberoamericano y es poco visible internacionalmente."

Finalmente el autor Burque-Gámez concluye: i) La información de la producción científica permitirá realizar un diagnóstico fiable de recursos científicos y humanos, y priorizar investigaciones; ii) Ecuador necesita un Plan Nacional de Ciencia y Tecnología que estimule la investigación para el cambio de la matriz productiva, y iii) las publicaciones de los ecuatorianos tienen bajo Factor de Impacto (en promedio están en 3)[4].

Por tanto, para concentrar los esfuerzos de los investigadores del país, es necesario que cada uno conozca a los autores nacionales y externos que tienen líneas de investigación afines y con quienes podría contactar con el objetivo de maximizar la contribución que en equipo se podría conseguir.

2.2. Uso de datos enlazados para la creación de redes de colaboración

En la actualidad, la aplicación de las guías de diseño para la publicación de datos enlazados en la Web (Berners-Lee, 2006) y tecnologías de la Web Semántica, con el objetivo de reducir barreras al momento de integrar o interoperar datos entre silos heterogéneos, ha concitado el interés de diferentes comunidades (gobierno, organizaciones educativas, investigadores, negocios, entre otros). La interacción entre datos ha permitido el descubrimiento de nueva información y conocimiento mediante la explotación de datos enlazados, que puede resultar en la solución a diversos problemas en diversas áreas de conocimiento. A continuación, se presentan algunos fundamentos teóricos en relación a estas tecnologías.

Linked data

Siguiendo los lineamientos de la W3C, la utilización de tecnologías de Linked Data permite vincular datos distribuidos en la Web. Esto contribuye a la Web Semántica mediante el proceso de publicación de datos: además de permitir que éstos puedan vincularse a otros datos, para enriquecerse y proporcionar las mejores formas de exploración para las personas y las máquinas mediante el uso de descripciones estándares en RDF (*Resource Description Framework*[5]).

La publicación de datos enlazados según la propuesta de (Berners-Lee, 2006), implica el seguimiento de los siguientes aspectos de diseño:

- Utilizar URIs como nombres para las cosas.
- Utilizar URIs desreferenciables para que la gente pueda ver esos nombres.
- Cuando alguien busca una URI, proporcionar información útil, utilizando los estándares apropiados (RDF para describir recursos y SPARQL para acceder y consultar datos).
- Incluir enlaces a otras URIs de tal manera que se pueda recuperar información.

Además, tomando en cuenta la filosofía de apertura de datos mediante licencias abiertas, Berners-Lee (2006) estableció cinco principios logrando así establecer una metáfora con estrellas según se avance en el nivel calidad de "*linked open data*", como sigue:

[4] http://www.telegrafo.com.ec/opinion/columnistas/item/la-produccion-cientifica-ecuatoriana-ii.html
[5] http://www.w3.org/RDF/

★	(1)	Disponible en la Web (en cualquier formato) pero con una licencia abierta.
★★	(2)	Disponible como datos estructurados legibles por máquina.
★★★	(3)	Igual que (2) pero en formatos no propietarios.
★★★★	(4)	Igual que (1), (2) y (3), además utilizar estándares W3C (RDF, SPARQL)

Para llevar a cabo la publicación de datos enlazados existen varias metodologías, que han evolucionado de acuerdo a las necesidades descubiertas en cada dominio de trabajo, además son prácticas sugeridas mas no impuestas para realizar el proceso de *linked data*. El uso de una metodología para la generación y publicación de datos enlazados permite obtener datos enlazados de calidad y mantener un orden a la hora de enlazar los datos a otras fuentes para enriquecer aún más la data generada.

Vocabularios Abiertos

Méndez & Greenberg (2012) en su trabajo destacan que la búsqueda por materias o conceptual, se ha vuelto la forma más utilizada de búsqueda en Web, esto implica que para satisfacer esta necesidad, necesitamos una búsqueda semántica. Los vocabularios controlados forman una parte fundamental en el proceso de datos enlazados ya que permitirán modelar determinado dominio basado en conceptos y relaciones.

Con el objetivo de maximizar las posibilidades de reuso de los datos Web, se deberían elegir vocabularios abiertos y consensuados. Si los diferentes trabajos académicos estuvieran descritos según estos esquemas estandarizados, las máquinas o agentes de software podrían procesar automáticamente la información de publicaciones con diferentes fines, en este caso para detectar redes de colaboración.

Para modelar el dominio de las publicaciones científicas se necesitó revisar varios vocabularios, utilizando el servicio de búsqueda *Linked Open Vocabularies*[6], un ecosistema de vocabularios abiertos enlazados para consulta y uso en los diferentes dominios, según sea el caso.

Algunas propuestas para representar publicaciones de varios tipos fueron encontradas, en este trabajo se optó por reutilizar algunas clases y propiedades de diferentes vocabularios relacionados al dominio de trabajo. En la Fig. 1 se muestra una de las aproximaciones desarrolladas para el dominio de las publicaciones científicas.

El dominio correspondiente a las publicaciones científicas se puede modelar reutilizando vocabularios como Bibo[7], Vivo[8], Dublin Core[9] y FOAF[10]. Los vocabularios Bibo y Vivo están diseñados para representar conceptos como: Artículo y Journal, así como propiedades que representa características de una publicación científica como por ejemplo DOI, volumen, fecha de publicación, entre otras. El vocabulario FOAF está diseñado para representar personas y relaciones entre éstas; permite modelar a personas y concretamente al Autor de la publicación y algunas propiedades que identifican datos personales como nombres, apellidos; y, la relación a sus afiliaciones (Organización). Finalmente, propiedades de Dublin Core permiten modelar propiedades como son el título de la publicación, abstract y fechas de publicación.

---

[6] Linked Open Vocabularies http://lov.okfn.org/dataset/lov/
[7] Bibliographic Ontology Specification, (2009), http://bibliontology.com/specification
[8] VIVO-ISF Ontology version 1.6, http://www.essepuntato.it/lode/owlapi/http://vivoweb.org/ontology/core# http://www.essepuntato.it/lode/owlapi/http://vivoweb.org/ontology/core#dataproperties
[9] Dublin Core Metadata Element Set, Version 1.1 , (2012), http://dublincore.org/documents/dces/
[10] FOAF Specification, http://xmlns.com/foaf/0.1/

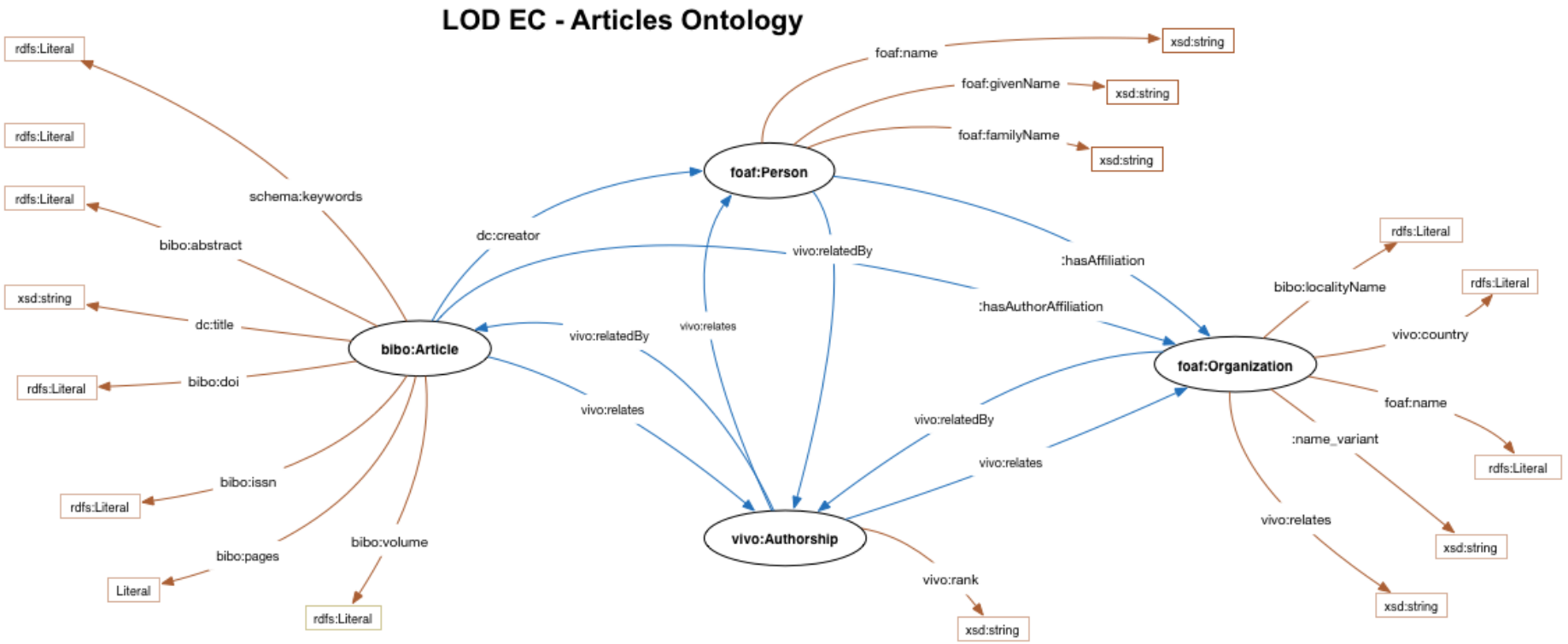

Figura 1. Vocabularios para la descripción de publicaciones científicas.

Sistemas de organización de conocimiento

Los Sistemas de Organización de Conocimiento (*Knowledge Organization System* o KOS por sus siglas en inglés), según Hodge (2008) -una de las primeras autoras en referirse a este término-, intentan abarcar a todos los tipos de esquemas para organizar la información y gestionar el conocimiento. Este tipo de sistemas, incluyen: i) esquemas de clasificación y categorización que organizan materiales a nivel general, ii) los encabezamientos de materia que proporcionan un acceso más detallado a los materiales, y iii) los ficheros de autoridades que controlan las versiones variantes de información clave como nombres geográficos y nombres personales. Hoy en día, los KOS también incluyen vocabularios altamente estructurados, como tesauros, y esquemas menos tradicionales, como las redes y ontologías semánticas (Hodge, 2008).

En grandes repositorios de información, como la Web, los sistemas de clasificación de conocimiento son utilizados por su capacidad para organizar información en relación a determinadas categorías de interés; esta característica facilita la gestión y localización de material.

En el contexto de la Web Semántica, se ha creado el vocabulario SKOS[1] (*Simple Knowledge Organization System*), estándar que a la presente fecha se ha convertido en una recomendación del W3C.

SKOS proporciona una forma estándar para representar un sistema KOS mediante el lenguaje RDF; es decir, permite describir esquemas de conceptos como tesauros, esquemas de clasificación, taxonomías y otro tipo de vocabulario controlado, garantizando así la interoperabilidad entre aplicaciones (Francesconi *et al*., 2008).

Actualmente, SKOS ha sido utilizado para representar conocidos tesauros de organización de material en bibliotecas o repositorios digitales, como:

- *Dewey Decimal Classification* (DDC[2])
- *Universal Decimal Classification* (UDC[3])
- *UNESCO Nomenclature*[4]
- *Joint Academic Coding System* (JACS[5])

Además de los vocabularios controlados que se han mencionado, el uso de SKOS se ha proliferando en fuentes de conocimiento basados en servicios sociales. De forma concreta, el estándar SKOS es utilizado para organizar categorías de conceptos publicados de forma abierta y colaborativa en uno de los servicios sociales más populares, la Wikipedia.

Si en la web social, la Wikipedia es uno de los principales casos de éxito, en la Web Semántica, la DBPedia[6] es la fuente de datos estructurados más reconocida.

La ontología DBPedia permite describir cualquier tipo de entidad del mundo real; el repositorio de DBPedia incluye datos RDF derivados de la Wikipedia. Además los recursos de DBpedia incluyen vínculos a entidades a otras fuentes como, YAGO, OpenCyc y WordNet, permitiendo así, consolidar e integrar información de un mismo sujeto (persona, institución, producto, localización, etc.), a través de diferentes fuentes heterogéneas.

En la Fig. 2, se muestra la representación de conceptos relacionados a la categoría DBpedia, Learning[7]. Como se puede observar, el predicado SKOS, broader, permite vincular un término a su concepto superior.

La ventaja de sistemas de datos abiertos como la DBPedia, es que se mantiene actualizado gracias a que la gente continuamente crea, edita y organiza los contenidos de la Wikipedia. Por tanto, a diferencia de los vocabularios cerrados, en la DBPedia es posible consultar información sobre nuevas áreas o tópicos de conocimiento.

---

1 http://www.w3.org/TR/skos-reference/
2 https://www.oclc.org/dewey.en.html
3 http://www.udcc.org/
4 http://databases.unesco.org/thesaurus/
5 http://www.hesa.ac.uk/content/view/1787/281/
6 http://dbpedia.org/
7 http://dbpedia.org/resource/Category:Learning

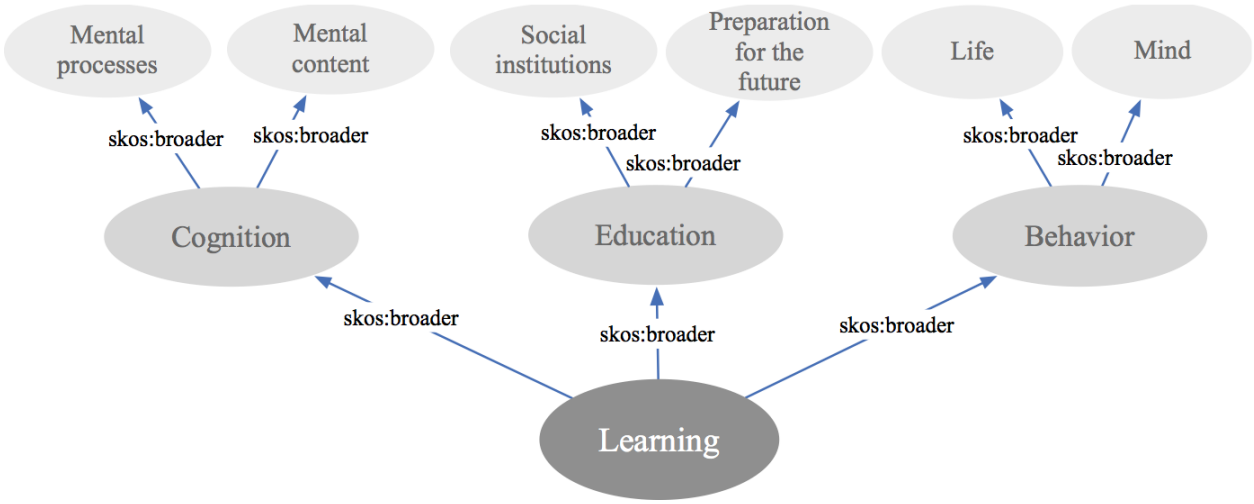

Figura 2. Taxonomía representada en SKOS: meta-conceptos relacionados a "Learning".

Para conseguir el propósito de este trabajo, se enriquecerán las palabras clave asociadas a las publicaciones, aprovechando las relaciones semánticas SKOS definidas entre términos y las miles de descripciones de todo tipo de tópicos que se encuentran disponibles en DBPedia.

2.3. Trabajos relacionados

En Ecuador, la aplicación de tecnologías semánticas y el uso de datos enlazados, aún es incipiente, por tanto, son buenas las posibilidades de promover la investigación y el desarrollo de esta área, enfocando los esfuerzos en resolver problemas que afectan a nuestro entorno y/o comunidad.

En el sitio de dataHub.io, plataforma de gestión de datos de la Open Knowledge Foundation, se encontraron 16 repositorios que mencionan a Ecuador[8]. De esta cantidad, dos *datasets* describen a instituciones del país (Universidad de Cuenca[9] y Universidad Técnica Particular de Loja[10]); un dataset proporciona acceso abierto a información sobre las publicaciones indexadas en SCOPUS con al menos un autor de afiliación ecuatoriana, "Open Data of Ecuador[11]"; finalmente, se encontraron las iniciativas: GeoEcuador[12] y Serendipity[13], en los ámbitos de información geoespacial y de recursos educativos abiertos, respectivamente. El resto de *datasets* incluyen referencias menores a Ecuador. Los casos mencionados, han surgido gracias a la colaboración de investigadores del sur del país, equipo del cual forman parte los autores del presente trabajo.

En el contexto internacional, el proyecto ReSIST[14], financiado en sus inicios por la Unión Europea, creó RKBExplorer, plataforma donde se pueden encontrar descripciones, en lenguajes de la Web Semántica, de publicaciones académicas. Entre otras bases indexadas de las que se puede explorar información, están: ACM, eprints, IEEE y DBLP.

En cuanto a propuestas para la detección de potenciales redes de colaboración, una de las aproximaciones más sencillas es explotar las relaciones de co-autoría de quienes trabajan en la publicación de trabajos científicos. Un ejemplo de este enfoque, se presenta en (Savić *et al.*, 2015), donde se aplicaron técnicas de detección de comunidades para construir una red de co-autores de trabajos publicados en un grupo de revistas de matemáticas de Serbia; la red representa la colaboración científica entre autores que publicaron sus trabajos en el período de 1932 a 2011.

---

[8] http://datahub.io/dataset?q=Ecuador
[9] http://datahub.io/dataset/universidad-de-cuenca-linkeddata
[10] http://datahub.io/dataset/utpl-lod
[11] http://datahub.io/dataset/opendataec
[12] http://datahub.io/dataset/geoecuador
[13] http://datahub.io/dataset/serendipity
[14] http://www.rkbexplorer.com/about/

Shin *et al.* (2008) presenten un enfoque diferente, en este caso, se trató de detectar campos de colaboración e interés entre investigadores, mediante un enfoque de redes sociales. Un trabajo similar corresponde a (Dias & Moita, 2014); en esta propuesta, se construyó una red de palabras clave, con el objetivo de identificar las términos más relevantes de un conjunto palabras y destacar los que tienen el mayor impacto en la red.

Para mejorar la capacidad de respuesta del sistema ante situaciones en las que puede no tener información, en ninguna de las propuestas que se han mencionado utilizan fuentes de datos abiertas, estructuradas e interoperables. En este trabajo, la identificación de potenciales comunidades científicas no se basa, únicamente, en la búsqueda de coincidencias exactas de palabras clave, sino que se consigue el enriquecimiento de *keywords* mediante una gran cantidad de conceptos disponibles en DPBedia.

3. MARCO DE TRABAJO BASADO EN TECNOLOGÍAS DE LA WEB SEMÁNTICA

En este apartado, se presenta el marco de trabajo para detectar potenciales redes de colaboración, a través del enriquecimiento semántico de las palabras clave asociadas a los trabajos científicos.

Mediante este trabajo, se intenta demostrar que los sistemas de organización de conocimiento sociales, publicados bajo principios de Linked Data, pueden ayudar a detectar potenciales comunidades científicas. En la Fig. 3 se puede observar que publicaciones y autores, en lugar de relacionarse directamente a través de palabras clave, se vinculan a través de conceptos SKOS. Como se demostrará más adelante, esta forma de organización de la red facilita encontrar pares de autores con intereses similares y que antes no hayan publicado juntos.

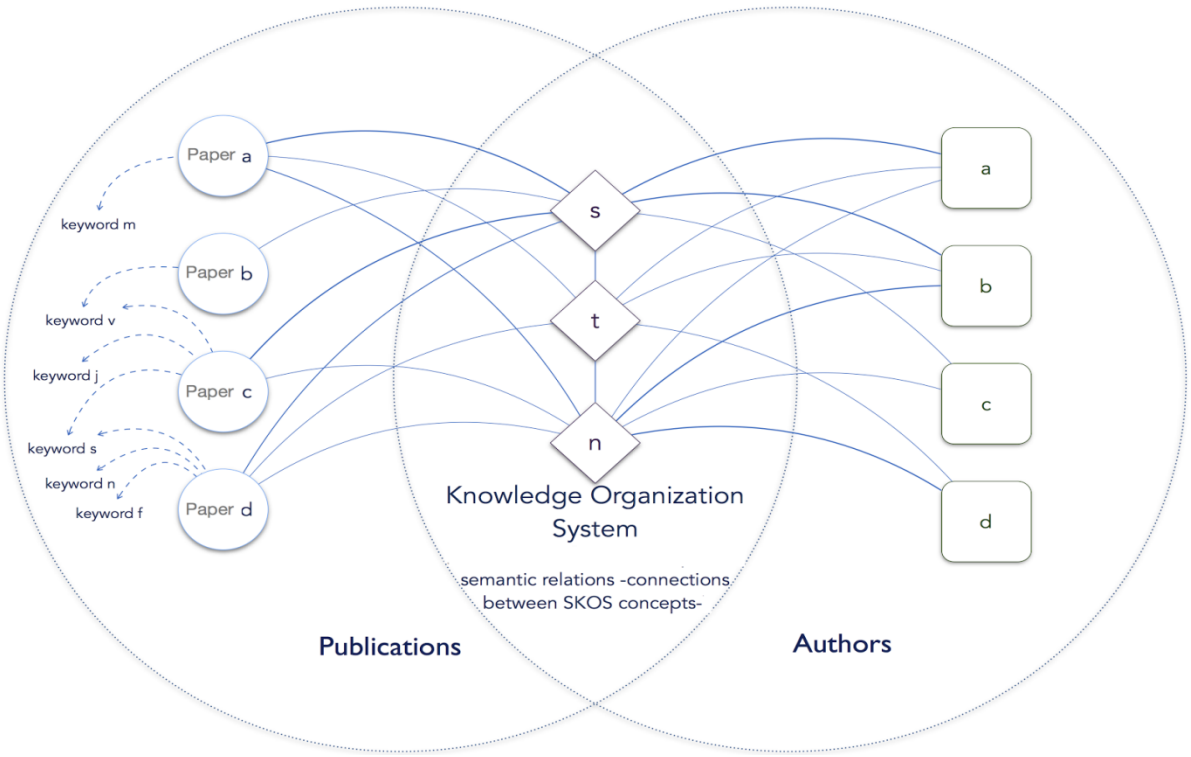

Figura 3. Sistema recomendador basado en conceptos SKOS.

Para llegar a construir una red de publicaciones es necesario que las publicaciones científicas, sus propiedades y relaciones con otras entidades, como autores e instituciones, estén expresadas en tecnologías semánticas. En este trabajo se ha seguido la metodología propuesta por los autores en trabajos anteriores (Piedra *et al.*, 2014); el *framework* comprende cinco fases:

- Identificación y selección de fuentes de datos, utilizando criterios de selección que depende de los datos que son relevantes, el tipo de licencias que poseen además del uso y usuarios de éstos.
- Modelamiento de vocabulario, que permite mapear conceptos a través del uso, reutilización o creación de vocabularios con la finalidad de obtener un modelo que defina el dominio de datos con el que se está trabajando. En esta fase también se puede llegar a determinar el modelo de las URIs a utilizar.
- Generación de datos en formato RDF, cuyo resultado son tripletas generadas previa limpieza, desambiguación y reconciliación de datos.
- Publicación de datos, que consiste en el almacenamiento en un repositorio adecuado.
- Consumo y visualización de datos, a través de un SPARQL Endpoint y diferentes aplicaciones que permitan hacer uso de los datos generados.

Una vez que las descripciones de los recursos son accesibles para las máquinas, es momento de construir aplicaciones y servicios que aprovechen la estructura y semántica de los datos y relaciones expresados en RDF. A continuación, se explica el proceso concreto para la detección de redes de colaboración, una de las aplicaciones que es factible implementar en la última capa del ciclo de publicación de datos enlazados.

En la Fig. 4, se presenta el proceso general propuesto para encontrar redes de colaboración aprovechando la estructura de grafo de redes construida con tecnologías semánticas y sistemas de organización de conocimiento como SKOS, (ver Fig. 3).

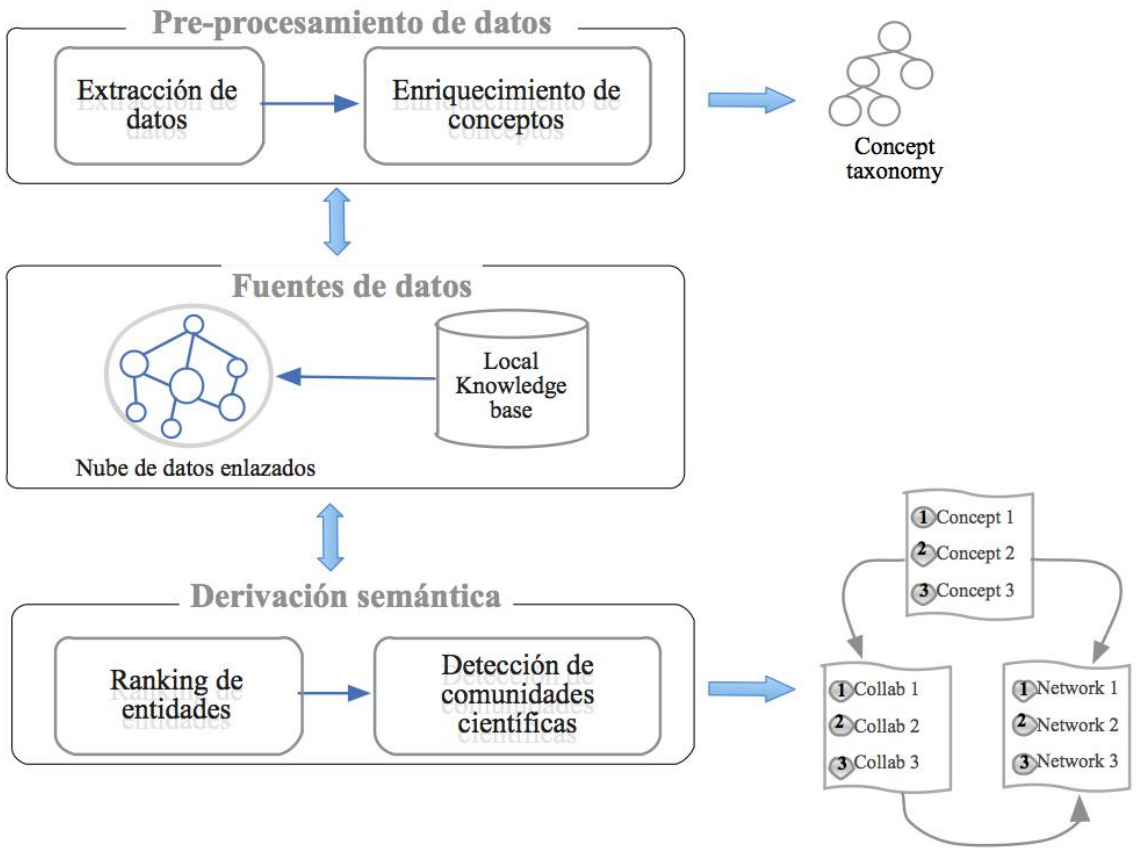

Figura 4. Proceso para la detección de redes de colaboración.

3.1. Pre-procesamiento de datos

El objetivo de este componente es crear una red de conceptos alrededor de cada palabra clave asociada a las publicaciones. Esta extensión del grafo de partida, permitirá realizar inferencias de conceptos relacionados al tópico para el cual se requiere encontrar redes de colaboración.

Esta capa inicia con la selección de fuentes de datos enlazados que describen publicaciones científicas. También incluye la identificación de repositorios de datos RDF a partir de los cuales se pueden enriquecer los tópicos.

Desde las fuentes seleccionadas, la tarea de extracción de datos puede ser realizada mediante servicios de consulta como SPARQL EndPoint. La información que sea recolectada puede ser almacenada en un repositorio local para reducir tiempos de procesamiento de las otras actividades.

Una vez que se tiene información sobre las publicaciones y se han creado enlaces entre los *keywords* del repositorio local y los correspondientes recursos en el *dataset* objetivo, se puede iniciar el proceso de enriquecimiento de cada palabra clave. El enriquecimiento consiste en encontrar una taxonomía de conceptos cercanos al nodo de interés; esto es posible, si se aplica un proceso iterativo de recorrido del grafo a través de las relaciones semánticas definidas por SKOS.

3.2. Derivación semántica

El propósito de este componente es detectar las comunidades o redes de colaboradores que han publicado juntos en relación a una temática.

La primera actividad que se debe realizar es agrupar autores que han publicado juntos en al menos un trabajo. A partir de la detección de las comunidades de colaboradores será posible recomendar a un autor determinado, los colegas que tienen intereses similares y con los cuales podría vincularse.

Como se puede ver en la Fig. 4, un aspecto clave en esta etapa, es determinar la relevancia de los conceptos y de los autores (ranking de entidades) que publican en relación a la taxonomía de conceptos generada en el componente de pre-procesamiento de datos. Aunque se pueden elegir diferentes criterios para determinar la relevancia de cada nodo, una aproximación básica es aprovechar la estructura de grafo de la red, y determinar así la cercanía de cada concepto a un objetivo dado; es decir, mientras más nodos sean necesarios visitar para llegar a un nodo dado, menos relevancia tendrá ese concepto, así:

$$Ranking\,(k) = \sum_{j=0}^{n} freq\,(c_j) * peso\,(c_j) \qquad (1)$$

donde:

- *cj* representa a cada concepto relacionado al keyword de interés (*k*).
- *freq(cj)* es la cantidad de veces que un concepto aparece en el grafo cercano a *k*, y
- *peso(cj)* es un número no mayor a 1, que es asignado a cada concepto de acuerdo al nivel de cercanía al nodo *k*.

La relevancia de un autor puede ser calculada a partir de las calificaciones de los conceptos que se relacionen con las publicaciones en las que ha participado.

4. RESULTADOS OBTENIDOS

A continuación, se plantea un escenario en el que se ha aplicado la presente propuesta. Como punto de partida, se obtuvieron datos sobre publicaciones indexadas en la base de datos SCOPUS. El dataset "*Open Data of Ecuador*[15]" proporcionó la información básica de cada publicación en la que aparece al menos un autor de afiliación de Ecuador. En total se encontraron 6468 trabajos académicos y científicos.

Para realizar una validación preliminar de la propuesta, se eligió un subconjunto de los trabajos, aquéllos publicados en revistas del área de Ciencias de la Computación. En la Tabla 1, se presenta un listado de las revistas o proceedings en las que se ha publicado 5 o más artículos y que son parte de los trabajos considerados en esta evaluación.

[15] http://datahub.io/dataset/opendataec

Tabla 1. Revistas con mayor cantidad de trabajos en el área de Ciencias de la Computación.

Publicación	Cantidad de trabajos
IEEE Latin America Transactions	8
Physica B: Condensed Matter	7
Expert Systems with Applications	6
IEEE Transactions on Learning Technologies	6
Information Sciences	6
Journal of Electrical and Computer Engineering	6
2010 IEEE Education Engineering Conference, EDUCON 2010	5
ICSTE 2010 - 2010 2nd International Conference on Software Technology and Engineering, Proceedings	5
IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans	5

Filtrando las publicaciones de Ecuador a aquéllas pertenecientes a Ciencias de la Computación, el corpus se redujo a 300 publicaciones. A través de una consulta al dataset "*Open Data of Ecuador*", se identificaron 746 keywords relacionados a esos trabajos. A partir de conjunto de datos se llevó a cabo el proceso descrito en la Fig. 3.

4.1. Pre-procesamiento de datos

La DBPedia fue elegida como fuente para el enriquecimiento de las palabras clave. Mediante el servicio de consulta SPARQL EndPoint[16] se recorrió el grafo de DBPedia, para encontrar los conceptos relacionados al término de interés. El uso del predicado skos:broader fue clave para poder obtener hasta tres niveles de relación.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT DISTINCT * WHERE
{
        <%s> <http://www.w3.org/2004/02/skos/core#broader> ?c1 .
        OPTIONAL {
                ?c1 <http://www.w3.org/2004/02/skos/core#broader> ?c2 .
        }
        OPTIONAL {
                ?c1 <http://www.w3.org/2004/02/skos/core#broader> ?c2 .
                ?c2 <http://www.w3.org/2004/02/skos/core#broader> ?c3 .
        }
}
```

Luego de ejecutar un script con la consulta indicada, se obtuvieron más de 6000 conceptos -o categorías- relacionados a los 746 keywords tomados como referencia. Explicando un caso concreto, para el término *Open Educational Resources* (OER[17]) se obtuvo una taxonomía como la que se indica: Education -> Educational Technology -> Educational materials -> Open Educational Resources.

[16] http://dbpedia.org/sparql

[17] http://dbpedia.org/resource/Open_educational_resources

4.2. Derivación semántica

En este componente, diferentes scripts basados en consultas fueron ejecutados para obtener: i) publicaciones relacionadas a los conceptos encontrados en la fase anterior, ii) autores de tales publicaciones, iii) conjunto total de palabras clave asociadas a esas publicaciones, y iv) metaconceptos a los que pertenecen las palabras clave. Al combinar esta información, se generaron los siguientes resultados para el caso del término: *Open Educational Resources*: de las 300 publicaciones relacionadas a Ciencias de la Computación, 18 mencionan a OER o a alguno de sus conceptos asociados; además 41 personas, que representan a 12 afiliaciones diferentes, constan como autores de estos trabajos. En la Tabla 2, se indica un extracto de los autores que obtuvieron mayor calificación, por sus publicaciones en esta área.

Tabla 2: Listado de autores que publican en relación a OERs.

Autor - Afiliación	Ranking
Edmundo Tovar Caro - Universidad Politécnica de Madrid	1
Nelson O. Piedra - Universidad Técnica Particular de Loja	2
…	
Miguel De J Ramírez Baltazar Carranza Itesm	
Carina Viteri - Escuela Politécnica del Ejército Mauricio Hincapié Oscar Hugues Salas	6
Audrey Romero Peláez - Universidad Técnica Particular de Loja Ismar Frango Silveira	7
Xavier Ochoa - Escuela Superior Politécnica del Litoral Ecuador Antonio Silva Sprock	8

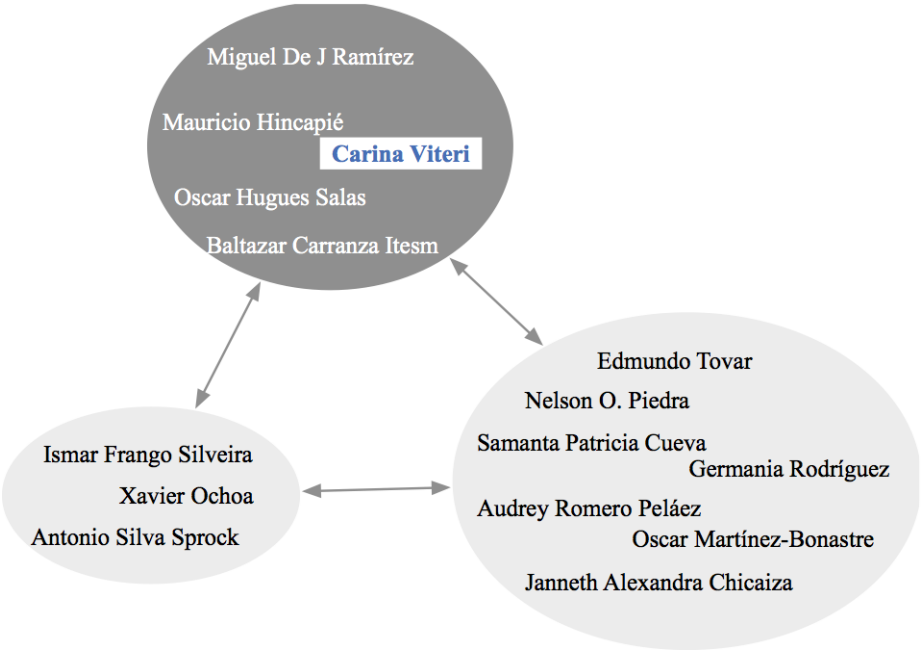

Figura 5. Comunidades de colaboradores.

La conformación de redes o comunidades de colaboración, se ejemplifica para un caso concreto. Para la investigadora, *Carina Viteri*, se ha encontrado la comunidad científica de la que es parte; su comunidad está constituida por todos los autores con quienes ha publicado al menos un trabajo indexado. Y se han encontrado al menos dos redes de colaboración de las cuales podría ser parte

puesto que registran trabajos en temas de OERs y conceptos asociados. En la Fig. 5, se muestran 3 redes de colaboradores y la comunicación que se podría establecer para conformar nuevas asociaciones.

5. CONCLUSIONES

La información que reside en bases de datos científicas debería ser visible y consultable a través de la Web. Un enfoque para conseguir este objetivo es publicar los metadatos como datos enlazados siguiendo los principios de Linked Data. Esto facilita que las propias máquinas tengan un nivel de comprensión de los recursos de información.

En el presente trabajo, se presentó un esquema para el análisis y enriquecimiento de metadatos de publicaciones científicas con el objetivo de detectar potenciales redes de colaboración. Para la anotación de artículos científicos se ha usado un modelo basado en el sistema SKOS, esquema utilizado en la base de datos enlazados más popular en la Web, la DBPedia. Las relaciones semánticas de SKOS permiten inferir vínculos jerárquicos que puedan ser utilizados para anotar la producción científica extraída. De esta forma es posible conectar entidades a través de las relaciones entre conceptos SKOS.

La propuesta fue validada considerando el caso de los artículos científicos publicados en la base de datos Scopus y con al menos un autor de afiliación ecuatoriana. El análisis permitió conocer algunas características de la estructura de las redes de colaboración tanto de los autores, como de las instituciones y de los países que participan, y a través de la anotación semántica los principales temas en los que trabajan los investigadores del país.

Las redes de colaboración detectadas son un insumo importante para fortalecer los esfuerzos del gobierno ecuatoriano y las autoridades universitarias del país, priorizar los esfuerzos y recursos invertidos en investigación y determinar la pertinencia o coherencia de los programas de investigación.

Los resultados preliminares, obtenidos en el escenario descrito en el apartado anterior, nos anima para seguir mejorando nuestra propuesta, y ponerla a prueba en escenarios con otras condiciones: conjunto de publicaciones, bases de datos o área de conocimiento diferentes.

REFERENCIAS

Berners-Lee, T., 2006. Linked data - Design issues. Disponible en *http://www.w3.org/DesignIssues/LinkedData.html*.

Burque-Gámez, S., 2013. La producción científica en Ecuador en el contexto Latinoamericano. Presentación realizada durante el seminario internacional: Herramientas para la difusión del conocimiento científico. Disponible en *http://www.senescyt.gob.ec/adjuntos/SEMINARIO_HERRAMIENTAS_CIENTIFICAS/5%20Sebastian%20Bruque%20Producion%20cientifica%20en%20Ecuador.pdf*.

Dias, T., G. Moita, 2014. *Identifying relevant keywords in scientific collaboration networks*. 11th World Congress on Computational Mechanics.

Francesconi, E., S. Faro, E. Marinai, G. Perugi, 2008. A methodological framework for thesaurus semantic interoperability. Proc. 5th European Semantic Web Conference, 76-87.

Hodge, G., 2008. *Systems of knowledge organization for digital libraries: Beyond traditional authority files*. The First Digital Library Federation Electronic Edition.

Méndez, E., J. Greenberg, 2012. Linked data for open vocabularies and HIVE's global framework. *El profesional de la información*, 21(3), 236-244.

Morshed, A., 2012. Role of vocabulary for semantic interoperability in enabling the linked open data publishing. *International Journal of Database Management Systems*, 4(5), 21-37.