Medios de Difusión Científica: una aproximación basada en Linked Data

Janneth Chicaiza[1], Lilia Quituisaca-Samaniego[2], Fabricio Montaño[1], Nelson Piedra[1], and Paúl Medina[3]

[1] Universidad Técnica Particular de Loja, Departamento de Ciencias de la Computación y Electrónica, 1101608 Loja, Ecuador
{jachicaiza,nopiedra,wfmontano}@utpl.edu.ec
http://www.utpl.edu.ec
[2] Numérica Investigación Innovación y Desarrollo, Quito, Ecuador
lilia.quituisaca.samaniego@gmail.com
http://www.numericaiid.com
[3] Universidad de la Fuerzas Armadas-ESPE, Departamento de Ciencias Exactas, Sangolquí, Ecuador
pmedinavz@gmail.com
http://www.espe.edu.ec
[2,3]Red Nacional de Investigación y Educación del Ecuador, Grupo de Trabajo de Repositorios Digitales, Cuenca, Ecuador
http://www.cedia.org.ec

Resumen La producción científica refleja los avances de una actividad académica o investigativa; además, también refleja el esfuerzo y los temas de vanguardia en los que se involucran las personas. Mejorar la visibilidad del trabajo que se realiza en cada campo de conocimiento es clave para propiciar el mejoramiento continuo de la investigación y, así, entre otras cosas, conseguir el reconocimiento de la comunidad científica internacional. A pesar de que existen diferentes medios para realizar la difusión de la producción científica, de manera particular las revistas (journals, proceedings, memorias, etc.) y los eventos académicos (congresos, meetings, workshops, etc.), no existe, hasta el momento, una plataforma que los consolide, discrimine y clasifique. Bajo esta premisa consideramos que si existiese una plataforma con las características señaladas, la información generada a partir de los distintos medios de difusión proporcionaría a los investigadores una fuente desde la cual podrían consultar y planificar, de acuerdo a cada temática o área, los potenciales lugares a los cuales podrían asistir o enviar sus trabajos. Para cumplir este cometido la presente investigación propone la creación de una plataforma que con el uso de tecnologías semánticas mejore la interoperabilidad de las fuentes de datos y, que basados en el paradigma de Linked Data, los datos de los distintos tópicos y recursos de carácter académico se vean enriquecidos por fuentes de conocimiento social, lo cual facilitaría la localización de resultados relevantes en la Web.

Keywords: Producción Científica, Divulgación, Recomendación, Linked Data, DBpedia, SKOS.

1. Introducción

La producción científica refleja los avances de una actividad académica o investigativa, en este sentido, también refleja el esfuerzo y los temas en los que se involucran las personas [1]. Por tanto, dar visibilidad al trabajo que se realiza en cada campo del conocimiento es clave para propiciar el mejoramiento continuo de la investigación y conseguir el reconocimiento de los investigadores.

La publicación de aportes o resultados de investigación en diferentes ámbitos de la ciencia permite alcanzar, entre otros los siguientes beneficios: i) los investigadores pueden compartir experiencias dentro de una comunidad y así acortar tiempos de investigación, continuar las líneas de investigación en común y proponer nuevas alternativas[4]; ii) las personas, grupos y organizaciones pueden ser evaluados de acuerdo a la calidad de su producción científica; y iii) la transmisión y difusión de los resultados de una investigación, en diferentes colectivos, puede producir potenciales beneficios para la comunidad y el ecosistema.

Con el objetivo de apoyar la búsqueda de los medios de difusión más adecuados, en el presente trabajo se aborda el diseo de una plataforma que consolida datos desde directorios o servicios donde se comparten convocatorias para presentación de trabajos y bases de datos científicas. A partir de esta información se generará un conjunto de potenciales convocatorias de acuerdo a una temática de interés. De esta manera, se espera apoyar al investigador en la identificación de los medios de difusión más adecuados para publicar sus trabajos, poniendo a su disposición potenciales canales de capacitación y difusión.

La propuesta ha sido diseada considerando un enfoque basado en tecnologías semánticas y Linked Data, es decir, aprovecha las estructuras semánticas definidas por ontologías, para generar nuevos hechos, y explota la gran cantidad de datos enlazados en la Web. Como resultado se generará un grafo, el cual será la base para realizar recomendaciones de los eventos y revistas de acuerdo a una temática determinada. Los tópicos de interés enriquecidos permitirán extender los resultados de una determinada búsqueda por medio de categorías relacionadas y características similares a los parámetros utilizados por el usuario durante su interacción inicial.

2. Marco Teórico

El fundamento de la Web es la construcción de un entorno global, abierto y extensible. Esta es la clave de su asombroso crecimiento y también la causa de nuevos desafíos que están pendientes de solución; especialmente los relacionados a la búsqueda, re-uso de información, personalización de contenidos y otros no menos importantes como, por ejemplo, la privacidad. Por otra parte, en la Web convencional, el contenido e información reside en fuentes o repositorios dispersos que no se comunican, por tanto, resulta complicado poder integrar, consultar y filtrar información.

[4] `http://www.utp.ac.pa/produccion-cientifica`

Actualmente, la Web Semántica provee la infraestructura y las tecnologías más prometedoras para integrar información. La Web Semántica [2] aade a la Web de Documentos, la semántica que le hace falta para disponer de un entorno en el que sea posible acceder a los datos contenidos en sitios Web y procesar automáticamente la información de un modo más exacto y completo.

Actualmente se está visionando la Web Semántica como una Web de Datos Enlazados; en esta Web las cosas se conectan a través de relaciones semánticas, a diferencia de la Web de Documentos. Además, mediante las tecnologías de Web Semántica, las máquinas pueden integrar, comparar y analizar datos codificados.

2.1. Linked Data

El paradigma de Linked Data ha evolucionado hasta convertirse en uno de los mayores retos en el ámbito de la gestión inteligente de la información: la explotación de la Web como una plataforma para la integración de datos y de información, así como para la búsqueda y consulta avanzada[3].

En la Web de Datos Enlazados, dos tecnologías son clave: i) el modelo de datos RDF, que permite describir cualquier entidad o recurso Web mediante un conjunto de predicados o relaciones; y ii) las ontologías o vocabularios abiertos que definen la estructura para describir los recursos.

En la Web de datos, cada recurso y cada predicado es identificado por una URI. En la Figura 1 se observa la representación semántica del país México, según DBPedia[5], popular repositorio de datos RDF que se alimenta a partir de las cajas de información de las páginas de Wikipedia.

En [4] se explica la estructura de los datos enlazados, los principios a los cuales se rigen y la forma de explotar estas estructuras mediante diferentes tipos de aplicaciones.

2.2. Sistemas de Organización de Conocimiento

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. En el contexto de la Web Semántica, se ha creado el vocabulario SKOS (*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 [5]. Los conceptos de un dominio de conocimiento codificados mediante RDF y descritos mediante SKOS permiten implementar sencillos mecanismos de inferencia que generen nuevas relaciones.

[5] `http://dbpedia.org/About`

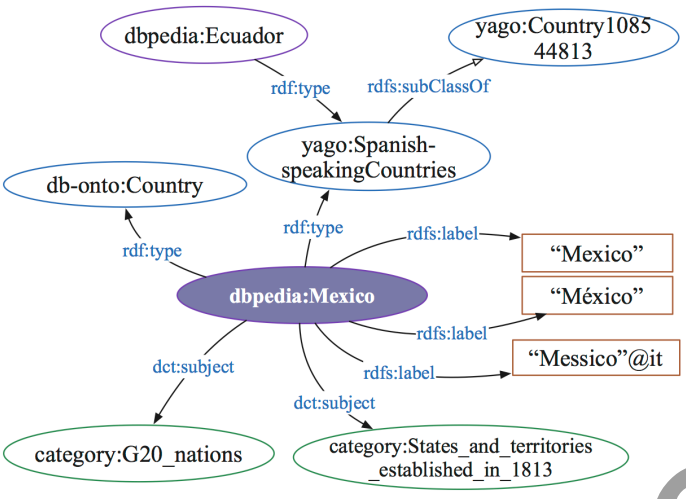

Figura 1. Grafo de la representación semántica del país México. Fuente DBPedia.

2.3. Trabajos relacionados

Los dos enfoques primarios para recomendar información, de acuerdo a las preferencias y contexto de un usuario son: i) los sistemas basados en contenido, y ii) los sistemas basados en colaboración. Uno de los principales problemas de los que adolecen estos sistemas es que requieren grandes cantidades de datos para poder ofrecer resultados precisos. Para mitigar este problema, actualmente se están implementando sistemas abiertos de recomendación basados en datos enlazados.

Por otra parte, los grafos RDF disponibles en la nube, ofrecen un catálogo de propiedades que puede enriquecer la información tanto de usuarios como de elementos a recomendar; por tanto, puede resultar útil un enfoque que complemente o mejore el rendimiento de los sistemas tradicionales.

A continuación se presentan algunas de las iniciativas que aprovechan el caudal de datos enlazados para ofrecer una experiencia personalizada al usuario.

- Passant en [6] propone uno de los primeros sistemas de recomendación basados en datos enlazados. En su trabajo describe la implementación de *dbrec*, un sistema de recomendación de bandas y artistas musicales. El motor de recomendación se basa en el algoritmo LDSD (*Linked Data Semantic Distance*) utilizado para calcular la distancia semántica entre dos recursos RDF.
- En [7] se experimenta con tres algoritmos diferentes para generar listados de los N primeros libros que el usuario posiblemente prefiera. Para generar las recomendaciones se trabajó con un dataset de libros calificados por usuarios.
- En [8] se propone un sistema recomendador colaborativo que aprovecha los nodos y enlaces de un grafo para mejorar la precisión y exhaustividad de un sistema cerrado. Mediante su propuesta, basada en un enfoque de filtrado

colaborativo, intentan afrontar 3 problemas puntuales: i) el problema del nuevo ítem; ii) el problema del nuevo usuario, y iii) la escasez de datos sobre ratings.

En este trabajo y a diferencia de los trabajos mencionados, se explotan jerarquías de conocimiento definidas por vocabularios como SKOS. En SKOS, los elementos de un tesauro son representados por medio de conceptos entre los que se establecen relaciones jerárquicas y, a partir de esta estructura es posible inferir un nuevo conocimiento que proporcione recomendaciones más acertadas.

3. Definición de la Propuesta basada en Linked Data

La implementación de la plataforma comprende la definición y construcción de componentes Web para la extracción e integración de datos desde diferentes fuentes y, el posterior tratamiento de esta información. La información procesada permitirá realizar recomendaciones más relevantes al momento de la búsqueda y presentación de las distintas opciones . En la Figura 2 se muestran los componentes y principales actividades que se deben desarrollar antes de tener una funcionalidad que pueda ser utilizada por los investigadores.

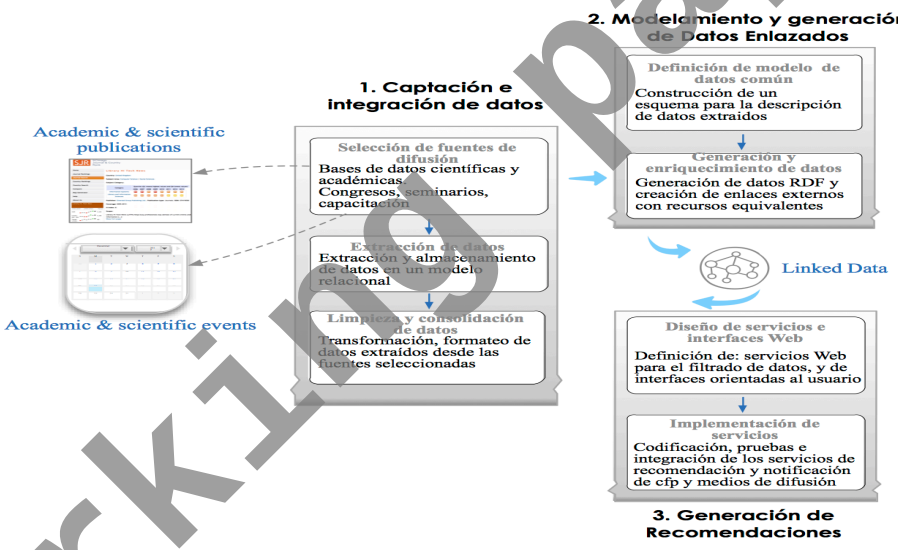

Figura 2. Metodología utilizada para la implementación del servicio. El ciclo inicia con la identificación y selección de catálogos de medios de difusión; luego, una vez que los datos han sido coleccionados, para mejorar la calidad de las recomendaciones se realiza el pre-procesamiento. A continuación, durante la fase de modelamiento y generación se pone a punto la información recolectada para que pueda ser consumida por los servicios de recomendación y notificación que forman parte de la tercera capa definida en la metodología.

Cabe indicar que como Linked Open Data (LOD2) se ha convertido en un marco de referencia para la construcción de aplicaciones basadas en enfoques de datos enlazados; ésta propuesta ha sido desarrollada de acuerdo a su ciclo de publicación, el cual se encuentra establecido en [9].

4. Implementación del Primer Prototipo

Como parte del proceso de creación del servicio de sugerencia de publicaciones, en esta sección se bosqueja la implementación del primer prototipo del servicio de localización de publicaciones.

4.1. Captación e integración de datos

En primer lugar se consideran bases de información científica; en particular, para esta primer prototipo se ha considerado a Web of Science, Scimago y Latindex. En segundo lugar, con el objetivo de encontrar catálogos de llamadas a presentación de trabajos se realizaron búsquedas en la Web, hallando sitios basados en herramientas sociales como por ejemplo, WikiCFP[6]. Una vez determinados los sitios se realizó una valoración de su "calidad", valoración que considero los siguientes criterios: orientación (áreas temáticas que cubre), frecuencia de actualización de datos, autoría, y factibilidad de compartición o extracción de datos; y, en tercer lugar, se buscaron repositorios de datos que se encuentren enlazados

A partir de las fuentes de datos seleccionadas, se crearon diferentes scripts para coleccionar los datos. De manera preliminar se utilizó la técnica de scrapy Web para leer datos desde páginas estructuradas y; por otra parte, se procesaron los archivos en formato .xls (Excel) proporcionados por las bases de datos científicas seleccionadas. Hasta el momento se ha recogido información del ranking de las revistas de diferentes aos, de diferentes disciplinas y, como se ha indicado, de diferentes sitios. Esta colección de información permite empezar con la fase de pre-procesamiento, analisis e integración de datos.

Para completar esta fase se analizaron los datos coleccionados y se detectó que algunos de ellos, en particular los que hacen referencia a lugares, temáticas y eventos presentaban diferencias léxicas según la fuente de procedencia; por tanto, se llevó a cabo una nueva fase de pre-procesamiento de datos con el objetivo de limpiar e integrar datos de objetos equivalentes. Estos datos formateados y unificados fueron asignados a los recursos originales.

4.2. Modelado y generación de datos

Una vez que se coleccionaron y pre-procesaron los metadatos de los catálogos seleccionados, se procedió a definir un modelo común para poder describir a todos los tipos de medios de difusión. El modelo común, basado en tecnologías de la Web Semántica, ha permitido implementar un servicio de recomendación que

[6] `http://www.wikicfp.com/cfp/`

acceda de forma integrada a la información recolectada, independientemente del tipo de recurso y de su estructura. El diseo de alto nivel del modelo conceptual para describir a los tipos de medios de difusión seleccionados puede observarse en la Figura 3.

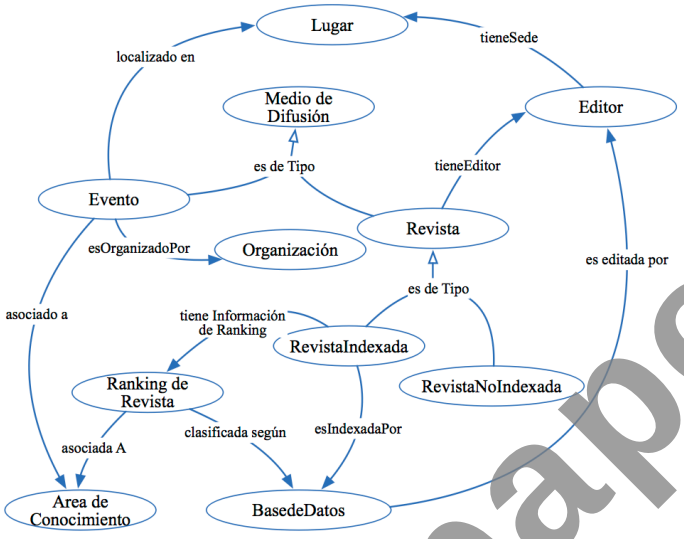

Figura 3. Modelo de alto nivel para describir medios de difusión

A partir del esquema de referencia definido, se generaron las tripletas y se procedió a enriquecer los datos de los medios de difusión seleccionados a partir de los datos disponibles en la nube de Linked Data. El enriquecimiento consiste en extender cada uno de los términos asociados a los medios de difusión y al usuario.

La extensión de términos comprende la inclusión de sinónimos, acrónimos, variaciones léxicas y conceptos relacionados (superiores o inferiores); esto último permite ampliar las posibilidades de encontrar material relevante, más allá de la clásica correspondencia léxica entre la consulta de usuario y el contenido de los recursos.

Las temáticas o áreas de conocimiento asociados a cada medio de difusión (eventos y revistas) han sido mapeados a un concepto DBpedia. La localización de términos relacionados a una determinada categoría de un evento o journal se realiza por medio del recorrido de la estructura jerárquica definida por el vocabulario SKOS.

4.3. Generación de Recomendaciones

La implementación del primer prototipo incluyó el diseo de servicios web REST (Representational State Transfer). Para acceder a los servicios se realiza

el llamado a través de una aplicación cliente (como un navegador web), con los parámetros necesarios para ser procesados por el servidor y generar un listado de recursos (revistas o eventos) en los que un usuario podría estar interesado. La codificación de los servicios fue realizada utilizando Flask[7] que es un micro-framework de Python.

Finalmente, para la construcción del cliente Web se han utilizado las tecnologías de HTML5, JavaScript y CSS3; de esta manera se facilita la interacción de los usuarios de una manera sencilla. La invocación a cada servicio se realiza a través de los parámetros de búsqueda.

5. Conclusiones

Existen diferentes medios para realizar la difusión de la producción científica; entre otras opciones, se pueden nombrar: revistas indexadas, revistas no indexadas, actas de congresos, informes técnicos, ensayos, etc. Así, encontrar las fuentes de difusión más pertinentes puede ser el primer paso para mejorar la visibilidad y el impacto de los resultados que en cada proyecto se generan; por lo tanto, este documento describe una propuesta para:

- desarrollar una infraestructura que permita optimizar los recursos académicos orientados a apoyar la difusión de los resultados obtenidos por los investigadores.
- implementar un servicio que soporte la recomendación de medios de difusión de acuerdo a los intereses de cada investigador. Aquí, en particular, se detectaron los siguientes hitos a cumplir:
  - definir un marco de trabajo para la implementación de la plataforma de recomendación,
  - seleccionar los catálogos de medios de difusión Web más adecuados y extraer los datos a partir de la información disponible,
  - mejorar la visibilidad y la localización de los medios de difusión científica; para ello se aplicará tecnologías Web Semántica y el ciclo de publicación de Linked Data y
  - generar y valorar las recomendaciones desarrolladas a partir de los datos enlazados disponibles en la Web.

Finalmente, la propuesta planteada pretende procesar las preferencias de los intereses de los usuarios no solo mediante palabras claves, sino que potenciar su efectividad de búsqueda a través de representaciones semánticas disponibles en la nube de Linked Data.

Agradecimiento

Este trabajo ha sido desarrollado con el apoyo de la Consorcio Ecuatoriano para el Desarrollo de Internet Avanzado (CEDIA) del Ecuador y el Grupo de Trabajo de Repositorios Digitales.

[7] `http://flask.pocoo.org/`

Referencias

1. Piedra, N., Chicaiza, J., Cadme, E., Guaya, R.: 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. Volume 5. Revista Maskana, Cuenca, Ecuador (2014)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5) (May 2001) 34–43
3. Auer, S.: Linked open data creating knowledge out of interlinked data. results of the lod2 project. Volume 8661. Springer-Verlag, Berlin, Heidelberg (2014) 1–17
4. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems **5**(3) (2009) 1 – 22
5. Francesconi, E., Faro, S., Marinai, E., Perugi, G.: A methodological framework for thesaurus semantic interoperability. In: Proceeding of the Fifth European Semantic Web Conference. (2008) 76–87
6. Passant, A.: dbrec a music recommendations using dbpedia. In: SWC'10 Proceedings of the 9th international semantic web conference on The semantic web, DERI, Springer-Verlag Berlin (2011) 209–224
7. Musto, C., Basile, P., Lops, P., de Gemmis, M., Semeraro, G.: Linked open data-enabled strategies for top-n recommendations. In: Proceedings of the 1st Workshop on New Trends in Content-based Recommender Systems co-located with the 8th ACM Conference on Recommender Systems, CBRecSys@RecSys 2014, Foster City, Silicon Valley, California, USA. (2014) 49–56
8. Heitmann, B., Hayes, C.: Using linked data to build open, collaborative recommender systems. In: AAAI Spring Symposium: Linked Data Meets Artificial Intelligence, 2010. (2010)
9. Mezaour, A.D., Nuffelen, B.V., Blaschke, C.: Linked open data creating knowledge out of interlinked data. results of the lod2 project. Volume 8661. Springer-Verlag, Berlin, Heidelberg (2014) 155–174