# Propuesta de Esquemas de Perfiles para la Recuperación de Datos Científicos para un Sistema de Recuperación de Información del Área de Ciencias de la Computación

Rey, M.[1], Kuna, H.[1], Martini, E.[1], Canteros, A.[1], Cantero, A.[1], Rambo, A.[1], Biale, C.[1]

[1]Departamento de Informática. Facultad de Ciencias Exactas, Químicas y Naturales.
Universidad Nacional de Misiones
hdkuna@unam.edu.ar; martinrey@fceqyn.unam.edu.ar

**Resumen.** Un Sistema de Recuperación de Información requiere de diversos componentes que definen su funcionalidad y comportamiento. En el caso de un meta-buscador para la recuperación de datos científicos del área de ciencias de la computación, un esquema que defina la forma en la que van a ser almacenados tales datos se considera como un elemento necesario para su evolución. Se han desarrollado perfiles unificados para el almacenamiento de los datos de las entidades involucradas en la gestión de datos científicos, generados a partir de la acción de publicar un documento científico. Tales perfiles, se consideran como el punto de partida para la generación de nuevos componentes para el meta-buscador que haciendo uso de información propia puedan producir información de relevancia para el usuario de la herramienta.

**Palabras clave**: datos científicos, meta-datos, meta-buscador, recuperación de información.

## 1 Introducción

Los datos científicos pueden describirse como aquellos que son generados a partir de la acción de publicar un documento científico, en esta acción la publicación queda ligada a su autor o autores, así como también al medio que haya sido seleccionado para su divulgación, entendiendo al mismo como la revista científica o congreso de una disciplina en particular, elemento que agrega un dato más al considerar el campo de estudio en el cual se publica el documento. Posteriormente, si el artículo es citado por otra publicación se generan datos de esta conexión, ligando documentos, autores y fuentes de publicación. Considerando la registración de estas acciones a lo largo del tiempo, se podría generar un historial de un autor, de sus publicaciones y de los campos de conocimiento en los que suele trabajar. Inclusive se podría estimar la relevancia de sus trabajos y también de revistas o eventos científicos, al obtener información histórica de artículos que hayan sido publicados en los mismos.

La recuperación de datos científicos como actividad dentro de la disciplina de recuperación de información ha cobrado mayor interés en el último tiempo [1, 2]. El impacto de internet y sus tecnologías relacionadas ha llevado a la generación de grandes conjuntos de datos derivados de las acciones antes mencionadas [3, 4]. Asimismo, estos datos propiciaron el desarrollo de herramientas para su gestión, mantenimiento,

publicación y procesamiento. Es así como diferentes editoriales y asociaciones con reconocimiento de parte de la comunidad científica han publicado repositorios de datos científicos a través de sitios web que constituyen herramientas de consulta para el usuario-investigador. Complementariamente, han surgido numerosas herramientas de este tipo con un mayor o menor número de características, con variaciones en el origen de los datos, su organización y el procesamiento que realizan sobre ellos [5, 6].

Es en este contexto, que se ha planteado como alternativa un meta-buscador [7] que opera sobre el área de ciencias de la computación, accede a diversas fuentes para recuperar datos científicos y los ordena utilizando un algoritmo que estima su impacto en la comunidad científica [8, 9]. En esta herramienta, la búsqueda de integración de nuevas funcionalidades y de mejora tanto de rendimiento como de eficiencia exige la definición de una estructura homogénea para el almacenamiento de los datos científicos recuperados. Además de la consideración de cuestiones ligadas a la tecnología a emplear, en un entorno operativo cada vez más relacionado con Big Data.

Es por esto que el problema a resolver en la presente línea de trabajo es la generación de tal estructura; ya que es el punto de partida necesario para el desarrollo de procesos que, haciendo uso de los datos, generen información relevante para la consulta de un usuario-investigador, siendo este un objetivo a resolver en un futuro cercano.

## 2 Antecedentes

### 2.1 Un meta-buscador para datos científicos de las ciencias de la computación

El Sistema de Recuperación de Información (SRI) antes mencionado constituye un software que debe evolucionar [10]. En este sentido, se considera que un paso siguiente en su desarrollo comprende al almacenamiento de los datos de las entidades que son recuperadas a partir de cada búsqueda. De esta manera permitirá generar soluciones complementarias que, sin perder de vista el aspecto de *meta*-buscador, puedan servir para implementar servicios que sean de utilidad para los usuarios, y así obtener resultados de mayor relevancia. Por ejemplo: recomendar artículos similares, recomendar expertos para el área temática de la consulta, expandir las consultas del usuario a partir de resultados previos, entre otros.

Por otra parte, contar con tales datos, también implica definir diversas cuestiones ligadas a la implementación como son: un esquema de almacenamiento, tecnología asociada que permita su carga y consulta con facilidad, velocidad y efectividad, y los cambios que sean requeridos en los procesos seguidos por el SRI para la correcta integración de estas operaciones de persistencia. Tales definiciones constituyen el punto de partida para el desarrollo de operaciones que hagan uso de los datos almacenados para generar información relativa a las consultas que sean ejecutadas por los usuarios a fin de mejorar su experiencia con el meta-buscador.

## 2.2 Actualidad en la gestión de grandes conjuntos de datos científicos

En la actualidad, como se ha mencionado, existen numerosas alternativas a la hora de buscar datos científicos.

Por un lado se encuentran las asociaciones de profesionales y editoriales o empresas relacionadas que publican sus librerías digitales, como son ACM, IEEE y la editorial Elsevier con su producto Scopus [4]. Asimismo, existen diversos repositorios que recuperan y compilan publicaciones de revistas y eventos científicos, por ejemplo: DOAJ y DBLP [6]. A ellos, se suman desarrollos de buscadores que brindan acceso a una inmensa cantidad de fuentes de publicación, artículos y datos relacionados a sus autores, no solo las versiones académicas de los productos tradicionales de Google o Microsoft [11], sino también desarrollos de equipos de investigación u instituciones educativas [5]. Además de otros repositorios o bases documentales que se podrían denominar "particulares" al ser propias de una Universidad, Organización o Centro de Investigación específico, tanto del ámbito nacional como internacional[1].

## 2.3 Formatos para registración de datos científicos

Los diferentes lineamientos o recomendaciones que imponen las fuentes de publicación a través de las convenciones de formato de artículos que utilizan, generan que una misma entidad, pueda ser registrada de diferentes maneras según el lugar en dónde se publique. Este es un problema conocido en el ambiente académico, y en especial en el de procesamiento de datos científico; ejemplos concretos del mismo son: ambigüedad en el nombre de autores y centros de investigación, confusión en la relación entre las siglas y los nombres de revistas y congresos, dificultades en el tratamiento de nombres con formatos no-occidentales, entre otros [12, 13].

Es por esto, que con el correr del tiempo las fuentes de publicación han optado por la generación y adopción de esquemas para la catalogación de su contenido, en algunos casos siguiendo un estándar del área de bibliotecología como son Dublin Core o MARC [14], y en otros desarrollando esquemas propios. Los buscadores, en general han utilizado esquemas específicos a fin de dar a conocer las mejores prácticas para que el contenido de los repositorios fuera indexado y mejor posicionado por los motores de búsqueda, tal es el caso de herramientas como Google Scholar, Microsoft Academic y CiteSeer [15]. En otras oportunidades, las herramientas utilizan bases de datos (BD) que se generan a partir del procesamiento de publicaciones de fuentes específicas, por lo que los meta-datos disponibles son aquellos impuestos por sus convenciones de publicación internas. Mientras que existen otros casos en los que la definición de los meta-datos a almacenar se definió en base a procesos parcial o totalmente automáticos para recuperación de datos desde otras fuentes, tal es el caso, por ejemplo de DBLP [6], AMiner [5] y otras soluciones similares. Existen otros conjuntos de herramientas que permiten la registración de datos de un autor y sus publicaciones, pero en ellas, cambia el actor principal de la registración, no siendo un tercero, sino que son los mismos autores o investigadores aquellos que construyen los registros. Entre este tipo de

[1] Por ejemplo los disponibles en: *repositoriosdigitales.mincyt.gob.ar* – Accedido: julio-2016.

soluciones, se destacan las que generan redes sociales de investigadores como ResearchGate[2] y aquellas utilizadas para crear catálogos de recursos[3].

En este contexto, no se puede afirmar que exista un esquema consolidado a nivel global para el registro de meta-datos de entidades como son los artículos, autores, fuentes de publicación, áreas temáticas y centros de investigación, siendo que diferentes soluciones para el procesado de datos científicos operan bajo diferentes lineamientos que no llegan a constituir estándares para tales actividades. Por tal motivo, para una herramienta como el SRI [10], se hace necesaria la definición de estructuras necesarias para la transformación y persistencia de tales datos a fin de posibilitar el desarrollo de soluciones complementarias que puedan brindar información de mayor relevancia para un usuario-investigador, siendo éste el objetivo principal del presente trabajo.

# 3 Generación de perfiles para almacenar datos científicos

## 3.1 Diseño de los perfiles

A fin de generar la estructura para la gestión interna de los datos científicos se procedió a examinar un conjunto de fuentes como buscadores y BD[4] de carácter científico, de cada una de ellas se obtuvo un listado de los meta-datos que utilizan para la registración de las diferentes entidades con las que operan. Sobre ese relevamiento inicial, se prosiguió con la revisión de los atributos utilizados por cada fuente, a fin de identificar qué compone para cada una de ellas el "perfil" de cada entidad, por ejemplo: cuáles referencian todas las publicaciones de un autor a través de su perfil, cuáles especifican el área temática a la que se corresponde una revista o evento científico, qué tipo de métricas mantienen sobre cada entidad, entre otros. De esta manera, se obtuvo con una descripción detallada de las características principales de cada fuente que permitió progresar con una instancia posterior de análisis.

A partir del análisis de los meta-datos de cada fuente se prosiguió con la extracción de aquellos atributos comunes y/o de similar significado, para la caracterización de cada entidad. Posteriormente, se seleccionaron aquellos atributos, que pudiendo no estar incluidos en la operación anterior, se consideraron de importancia para los objetivos del SRI, y por lo tanto, se incorporaron a los perfiles a generar. Una vez definidos los perfiles, comenzaron a establecerse las relaciones entre los mismos, a fin de representar las conexiones lógicas que existen entre las entidades a registrar.

Como resultado de estas acciones, se obtuvieron los perfiles de meta-datos (ver figura 1) con los cuales operaría el meta-buscador, de las siguientes entidades:

- Artículos
- Autores
- Fuentes de publicación

[2] Sitio web: *researchgate.net/* - Accedido: julio-2016.

[3] Por ejemplo: Mendeley (*mendeley.com/*) – Zotero (*zotero.org/*) - Accedidos: julio-2016.

[4] Por ejemplo: ACM Digital Library, IEEE Xplore Digital Library, Google Scholar, Microsoft Academic, DBLP, CiteseerX, DOAJ, entre otros. - Accedidos: julio-2016.

- Áreas temáticas
- Centros de investigación

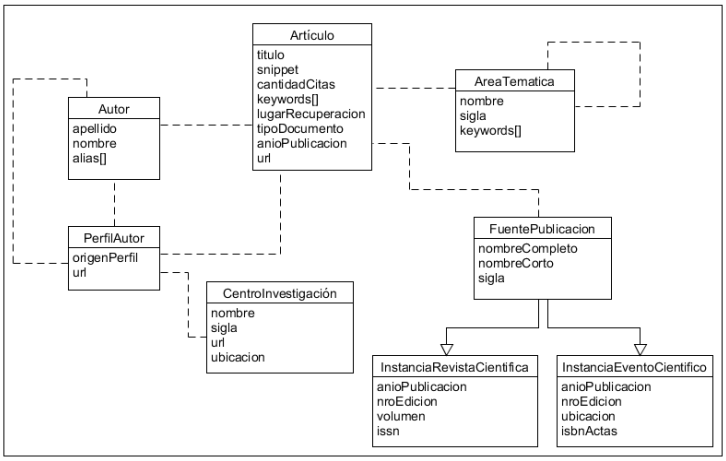


**Fig. 1.** Modelo conceptual de los perfiles generados para las entidades del SRI.

En las tablas 1, 2 y 3 se pueden apreciar la descripción de los atributos de las entidades principales del modelo: artículo, autor y perfil de autor y fuente de publicación con sus especializaciones para revistas y eventos científicos.

**Tabla 1.** Descripción de los atributos de la entidad Artículo

| Atributo | Descripción |
|---|---|
| Titulo | Título del documento |
| Snippet | Descripción del documento |
| CantidadCitas | Número de citas según la fuente consultada |
| Keywords[] | Términos clave asociados |
| LugarRecuperación | Fuente desde donde se recuperó el artículo |
| TipoDocumento | Identificación si es artículo, libro o pre-print |
| AnioPublicacion | Año de publicación del documento |

**Tabla 2.** Descripción de los atributos de las entidades Autor y Perfil Autor

| Atributo | Descripción |
|---|---|
| Apellido | Apellido del autor |
| Nombre | Nombres del autor |
| Alias[] | Conjunto de nombres recuperados del autor |
| OrigenPerfil | Fuente desde donde se recuperó el perfil |
| URL | URL del perfil |

**Tabla 3.** Descripción de los atributos de la entidad Fuente Publicación e Instancias

| Atributo | Descripción |
|---|---|
| NombreCompleto | Nombre sin abreviaturas de la fuente |
| NombreCorto | Nombre con abreviaturas |
| Sigla | Sigla de la fuente |
| AnioPublicación | Año en el que se publicó la revista o realizó el evento |
| NroEdición | Número de la edición de la revista / evento |
| Volumen / Ubicación | Identificación del volumen de la revista / Lugar de realización del evento |
| ISSN / ISBN | Identificador de la publicación |

### 3.2 Implementación de los perfiles

Una vez finalizado el diseño de los perfiles a emplear para la gestión de los meta-datos de las entidades con las que opera el SRI se procedió con su implementación en el back-end del meta-buscador. Para ello se desarrollaron los objetos correspondientes a tales entidades, respetando los perfiles antes descriptos e integrando a tales elementos dentro de la estructura del SRI, la cual se debió modificar para integrar nuevos módulos y generalizar el uso de los perfiles a través de todo el proceso de búsqueda y presentación de los resultados. De esta manera, la estructura del meta-buscador que fuera presentada previamente se amplió con los siguientes módulos:

- *Módulo para la gestión de la persistencia*: encargado de la configuración y ejecución de las operaciones cuyo objetivo es la interacción con la BD del SRI para almacenar y recuperar información de los perfiles.
- *Módulo para la recuperación de datos*: encargado del tratamiento y la actualización de los meta-datos de las diversas entidades que se almacenan en la BD del meta-buscador.

Las clases definidas para la representación de los perfiles de las entidades a gestionar se definieron en un ámbito general, con la finalidad de que todos los módulos de la herramienta pudieran hacer uso de los mismos en forma independiente al proceso que tengan por objetivo ejecutar. A nivel de tecnologías empleadas, el cambio con la última versión del SRI se presenta en la persistencia, migrando de un entorno relacional a uno

NoSQL[5], específicamente uno documental basado en objetos JSON[6] para el almacenamiento de las consultas ejecutadas y las entidades cuyos perfiles sean recuperados por los procesos del SRI. De esta manera, se integraron los componentes para la gestión de los perfiles al SRI, definiendo como siguiente paso la adaptación de sus procesos internos para hacer uso de los mismos a fin de brindar alguna solución complementaria para el usuario. La estructura del meta-buscador se puede observar en la figura 2.

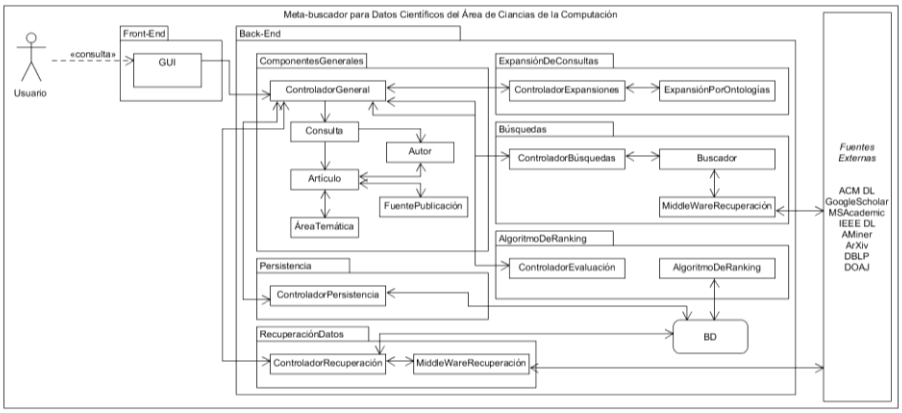


**Fig. 2.** Estructura del SRI

# 4 Implementación de los cambios en el SRI

## 4.1 Procesos del SRI modificados

Una vez integrados los perfiles definidos para la registración de los datos de las entidades con las que opera el meta-buscador, se debió proceder con la modificación de sus procesos operativos a fin de adaptar la solución desarrollada. Se inició con la modificación de los componentes del módulo de búsqueda de resultados, en los que se hace uso de las clases definidas por entidad para construir las colecciones de datos a procesar con cada búsqueda, almacenada en un objeto consulta en la BD. Por el momento, se mantuvo a las publicaciones científicas como único tipo de resultado a recuperar, sin embargo se enriquecieron los datos que se extraen de cada fuente con respecto a las relaciones que existen entre las entidades conforme a los perfiles del SRI.

A fin de optimizar los tiempos de respuesta al usuario, los meta-datos de los autores y las fuentes de publicación no son recuperados al mismo tiempo que se ejecutan las operaciones de búsqueda, sino que únicamente se recuperan los links hacia los perfiles que mantiene cada fuente consultada por el SRI, dejando rastro de las relaciones que existen entre tales elementos para la posterior recuperación de sus datos. De esta

[5] Acrónimo para Not Only SQL.
[6] Sigla para JavaScript Object Notation.

manera, la búsqueda se centra en recuperar los artículos necesarios para la aplicación del algoritmo de ranking y su presentación al usuario.

Para la ejecución de las operaciones posteriores a la búsqueda se han desarrollado procesos en los que no interviene el usuario y que tienen por objetivo la captura de los meta-datos de cada entidad que se haya generado en la ejecución del SRI, variando los detalles de transformaciones aplicables sobre los mismos a partir de las particularidades de cada fuente de datos que es consultada al momento de las búsquedas. De esta manera, se propuso un esquema a través del cual el SRI aumenta el contenido de su BD interna sin que esto signifique una merma en el rendimiento de las operaciones de búsqueda que son visibles al usuario-investigador.

### 4.2 Implementación de soluciones complementarias

Como resultado de las modificaciones introducidas, a partir de cada búsqueda la BD interna del SRI incorpora datos de los artículos que son recuperados, así como también de las entidades que guardan relación con los mismos. Por ejemplo: el perfil de un autor en particular según una o más fuentes, incluyendo sus artículos publicados, las áreas temáticas en las que ha trabajado, las instituciones a las que ha representado y los investigadores con quienes ha desarrollado relaciones de co-autoría. Al contar con este potencial volumen de datos se propuso el desarrollo de soluciones anexas que, integradas al SRI y haciendo uso de su BD, pudieran presentar información de valor agregado al usuario. Se plantearon tres alternativas, que se describen a continuación:

- *Método de recomendación de entidades basado en la relación que las mismas pudieran mantener con la consulta ingresada por el usuario*. Se propuso que a medida que el SRI ejecuta las búsquedas sobre fuentes externas, se presenten al usuario resultados recuperados desde la BD interna del meta-buscador, concretamente perfiles de autores, fuentes de publicación y artículos relevantes para la consulta ingresada por el usuario.
- *Método de filtrado de resultados basado en áreas temáticas*. En este caso, se propuso el desarrollo de un método que, considerando la consulta del usuario determine el área temática dentro de las ciencias de la computación con la que guarda mayor relación para poder aplicar filtros específicos al momento de que el SRI ejecute la recuperación de resultados desde fuentes externas.
- *Método alternativo de expansión de consultas basado en la interpretación de la consulta del usuario*. Se propuso que la consulta ingresada por el usuario pudiera ser expandida utilizando el "historial" de búsquedas del SRI.

## 5 Experimentación

### 5.1 Validación de los perfiles integrados al SRI

A fin de validar el funcionamiento del SRI al integrar los perfiles generados, se planteó un conjunto de pruebas que permitiera verificar el resultado de lo planteado en la sección 4.1. Para ello, se definió un conjunto de consultas a ejecutar en temáticas

variadas para evaluar que la captura, adaptación y persistencia de los datos fuera correcta. Con respecto al entorno tecnológico, los cambios con las últimas versiones del SRI se registraron en el backend, principalmente en la BD a emplear, migrando a un motor NoSQL de tipo documental, específicamente MongoDB en su versión 3.0.2.

Se realizaron 10 consultas, con un límite de 30 artículos a recuperar por cada una, ejecutando los procesos de recuperación de datos mencionados en la sección 4.1. Se evaluó la efectividad del SRI en tres aspectos: en el almacenamiento de los links hacia los perfiles, en la recuperación de los datos de la entidad y en el registro de las relaciones entre las entidades recuperadas a partir de un mismo documento. Los resultados obtenidos se pueden observar en la tabla 4.

**Tabla 4.** Resultados de la validación

| Métrica | Valor |
|---|---|
| Cantidad de perfiles a generar | 804 |
| Efectividad de la persistencia | **91%** (736/804) |
| Efectividad de la recuperación | **79%** (581/736) |
| Efectividad de la generación de relaciones entre entidades | **89%** (568/640) |

Al analizar los resultados obtenidos se puede observar que los métodos han funcionado en la mayoría de las ocasiones de manera correcta, con algunos errores en la recuperación de los perfiles, que deberán refinarse. Sin embargo, como el objetivo de la validación fue testear la funcionalidad e integración de los perfiles y procesos asociados, no se evaluaron a los mismos a partir de métricas propias de recuperación de información. Esta acción será ejecutada una vez implementados los procesos mencionados en la sección 4.2 cuando los datos de la BD del SRI sean explotados para la generación de información de ayuda para el usuario final.

## 6 Conclusiones y Líneas futuras de investigación

Se ha logrado especificar un conjunto de meta-datos para la registración de datos científicos. Los perfiles generados han sido producto de la unificación de los esquemas de almacenamiento planteados por diversas fuentes de datos, homogeneizando datos y relaciones a registrar. Estos perfiles se han implementado e integrado en un meta-buscador conjuntamente a los procesos necesarios para su generación, la recuperación de sus datos y establecimiento de las relaciones antes mencionadas.

A partir de la experimentación realizada se puede considerar que los mismos constituyen un punto de partida óptimo para el desarrollo de procesos destinados a generen información complementaria para optimizar los procesos del SRI o sean dirigidas al usuario para facilitar sus búsquedas. Tales soluciones son las principales líneas de trabajo a futuro a ejecutar, entre las que se destacan: un proceso de adaptación de consultas basado en algoritmos de extracción de tópicos y asociación por similitud y métodos para recomendación de resultados en forma interna basados en los datos presentes en la BD, entre otros. La implementación de este tipo de procesos se estima que implicará realizar adaptaciones sobre el meta-buscador, principalmente en un

aspecto técnico, requiriendo la utilización de estrategias propias del área de Big Data en cuanto al almacenamiento y acceso a datos.

# 7 Bibliografía


1. Bose, R., Frew, J.: Lineage retrieval for scientific data processing: a survey. ACM Computing. Surveys. CSUR. 37, 1–28 (2005).
2. Simmhan, Y.L., Plale, B., Gannon, D.: A Survey of Data Provenance in e-Science. SIGMOD Rec. 34, 31–36 (2005).
3. Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., Terliesner, J.: Coverage and adoption of altmetrics sources in the bibliometric community. Scientometrics. 101, 1145–1163 (2014).
4. Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., Pappas, G.: Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J. 22, 338–342 (2008).
5. Tang, J.: AMiner: Mining Deep Knowledge from Big Scholar Data. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 373–373. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2016).
6. Ley, M.: The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In: Laender, A.H.F. and Oliveira, A.L. (eds.) String Processing and Information Retrieval. pp. 1–10. Springer Berlin Heidelberg (2002).
7. Kuna, H., Rey, M., Martini, E., Solonezen, L., Podkowa, L.: Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación. Rev. Latinoam. Ing. Softw. 2, 107–114 (2013).
8. Kuna, H., Martini, E., Rey, M.: Evolution of a Ranking Algorithm for Scientific Documents in the Computer Science Area. In: XX Argentine Congress of Computer Science Selected Papers. pp. 145–155. EDULP, La Plata, Buenos Aires, Argentina (2015).
9. Rey, M., Kuna, H.D., Martini, E., Podkowa, L., Pautsch, J.G.A., Zamudio, E.: Generación de un método de expansión de consultas basado en ontologías para un sistema de recuperación de información. Presented at the XX Congreso Argentino de Ciencias de la Computación (Buenos Aires, 2014) (2014).
10. Kuna, H., Rey, M., Martini, E., Canteros, A., Cantero, A., Rambo, A., Biale, C., Zamudio, E.: Avances en la Construcción de un Sistema de Recuperación de Información para Información Científica en Ciencias de la Computación. Presented at the XVIII Workshop de Investigadores en Ciencias de la Computación April 14 (2016).
11. Ortega, J.L., Aguillo, I.F.: Microsoft academic search and Google scholar citations: Comparative analysis of author profiles. J. Assoc. Inf. Sci. Technol. 65, 1149–1156 (2014).
12. Ruiz-Pérez, R., López-Cózar, E.D., Jiménez-Contreras, E.: Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies. J. Med. Libr. Assoc. 90, 411–430 (2002).
13. Garciarena Ucelay, M.J., Villegas, M.P., Cagnina, L., Errecalde, M.L.: Cross domain author profiling task in spanish language: an experimental study. J. Comput. Sci. Technol. 15, no. 2, (2015).
14. Ortiz-Repiso Jiménez, V.: Nuevas perspectivas para la catalogación: Metadatos Versus Marc. Rev. Esp. Doc. Científica. 22, 198–219 (1999).
15. Beel, J., Gipp, B., Wilde, E.: Academic Search Engine Optimization (ASEO). J. Sch. Publ. 41, 176–190 (2009).