PA1830-5

Repositorio de conocimiento científico integrado conforme a *Linked Data*

Hernán Julián Gavilán Acosta

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
2018

PA1830-5

Repositorio de conocimiento científico integrado conforme a *Linked Data*

Autor:

Hernán Julián Gavilán Acosta

MEMORIA DEL TRABAJO DE GRADO REALIZADO PARA CUMPLIR UNO DE LOS REQUISITOS PARA OPTAR AL TITULO DE

MAGÍSTER EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Director

Luis Manuel Vilches Blázquez

Comité de Evaluación del Trabajo de Grado

Regina Motz

Víctor Saquicela

Página web del Trabajo de Grado

http://pegasus.javeriana.edu.co/~PA1830-5-LinkedRepo/

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERIA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
Noviembre,2018

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Rector Magnífico

Jorge Humberto Peláez, S.J.

Decano Facultad de Ingeniería

Ingeniero Jorge Luis Sánchez Téllez

Director Maestría en Ingeniería de Sistemas y Computación

Ingeniera Ángela Carrillo Ramos

Director Departamento de Ingeniería de Sistemas

Ingeniero Efraín Ortiz Pabón

Artículo 23 de la Resolución No. 1 de Junio de 1946

"La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia"

AGRADECIMIENTOS

Agradezco a mi familia que siempre ha sido mi apoyo en todos los proyectos que he emprendido en todas las formas en las que se puede apoyar a una persona. Han sido siempre la inspiración que me guía en la búsqueda del conocimiento y crecimiento personal, sin la influencia e impacto que han tenido en mi vida, no habría podido alcanzar este logro.

También quiero agradecer a todos los profesores que han compartido sus conocimientos conmigo y, especialmente, a Luis Manuel Vilches gracias a su disposición y guía fue posible asumir este reto y superarlo de forma exitosa.

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

ABSTRACT

This project presents a solution for the recovery, cleaning and structuring of data on publications of researchers from the Faculty of Engineering of the University Javeriana, specifically of the departments of Systems Engineering, Electronic, Civil, and industrial, according to the principles of Linked Data. Thus, this project allows to achieve an organization of the existing knowledge about said publications, Linked Data achieving the integration of distributed data in multiple and heterogeneous repositories, overcoming the obstacles of access and current management of these resources and alleviating the efforts on the part of the users of these data, such as researchers, analysts and directors of institutions related to scientific research.

RESUMEN

Este proyecto presenta una solución para la recuperación, limpieza y estructuración de datos sobre publicaciones de investigadores de la Facultad de Ingeniería de la Universidad Javeriana, concretamente de los departamentos de ingeniería de Sistemas, Industrial, Civil y Electrónica, conforme a los principios de *Linked Data*. Así, este trabajo permite conseguir una organización del conocimiento existente sobre dichas publicaciones, logrando la integración de datos distribuidos en múltiples y heterogéneos repositorios, superando los obstáculos de acceso y gestión actuales de estos recursos y aliviando los esfuerzos por parte de los usuarios de estos datos, tales como investigadores, analistas y directivos de instituciones relacionados con la investigación científica.

RESUMEN EJECUTIVO

Este trabajo se centra en los problemas presentes en el contexto del análisis bibliográfico de publicaciones para diferentes tipos de usuarios. La aproximación más frecuente consiste en emplear herramientas que consumen servicios disponibles en la red, alimentados por fuentes bibliográficas bajo demanda, que generan textos planos y permiten el análisis de las relaciones, pero no llegan a presentar una solución de fondo a los problemas de integración de datos, razón por la que se requiere consultar diversas fuentes y la revisión de resultados de forma manual, esto debido a que los datos bibliográficos se encuentran distribuidos en fuentes heterogéneas y desconectadas, por lo que, generalmente, se requiere de conocimientos técnicos para realizar análisis de patrones entre autores y temas de investigación empleando herramientas específicas. Las herramientas actuales no proporcionan una solución de fondo para la interconectividad de las bases de datos, ya que los repositorios que se encuentran disponibles en Internet son parte de iniciativas asiladas que no siguen un estándar global, aplican sus propias metodologías para la recuperación de datos y exposición de estos.

Los problemas enunciados afectan el acceso a información contenida en repositorios de publicaciones científicas y sus datos, que representa el primer paso a seguir para todo investigador previo al inicio de un proyecto. Para mejorar el proceso de recopilación de conocimiento, se genera un repositorio de conocimiento de publicaciones científicas, recopilando datos sobre publicaciones realizadas por parte de los profesores de la Universidad Javeriana, concretamente, de los departamentos de Sistemas, Industrial, Civil y Electrónica, clasificándolos e identificando su relevancia, para su tratamiento y publicación acorde a los principios de *Linked Open Data*.

Así, el foco de interés de este trabajo se encuentra en los datos disponibles en la web sobre publicaciones de investigadores de los mencionados departamentos de la Universidad Javeriana, a los cuales se accede mediante diferentes *Application Programing Interface* (API). Para la recuperación y gestión de los datos se desarrollan una serie de procesos de extracción, transformación y carga (ETL) de datos que permiten la ejecución de las fases de la metodología adoptada para la generación y publicación de *Linked Open Data*, empleando el *software* Pentaho y diversos lenguajes de programación (*Javascript* y *Python*).

Tras la recuperación de los datos, se realiza un proceso de modelado y descripción de los mismo, empleando modelos semánticos que permiten su posterior transformación a RDF, logrando así una integración semántica de los diversos repositorios considerados. De esta manera, se consigue que las publicaciones de la comunidad de la Facultad de Ingeniería se encuentren disponibles en una base de conocimiento, donde estas aparecen organizadas, estandarizadas e integradas semánticamente facilitando el acceso a los usuarios.

La generación y despliegue de esta base de conocimiento permite que la información pueda ser explotada consiguiendo revelar características sobre las publicaciones, autores y sus correlaciones que, de otra manera, permanecerían ocultas. Esto es posible gracias a la exposición y mantenimiento de un repositorio conforme a *Linked Data* que puede ser explotado y extendido, y resulta una fuente sobre la que es posible realizar desarrollos para la recuperación, tratamiento y analítica de datos.

INTRODUCCIÓN

Cada año un gran volumen de artículos científicos es generado y publicado en diversos repositorios de publicaciones disponibles para la comunidad, tales como: Scopus[1], Google Scholar[2], Elsevier[3], ResearchGate[4], etc. Estos repositorios, con frecuencia, ofrecen la información relacionada con dichas publicaciones a través de API (*Application Programming Interfaces*), en formatos como JSON (*JavaScript Object Notation*), XML (*eXtensible Markup Language*) o texto plano. Dichos repositorios contienen un enorme potencial para el trabajo científico, colaborativo, así como para la generación de nuevo conocimiento. Sin embargo, se encuentran dispersos y caracterizados por presentar una baja o nula conectividad entre los diferentes repositorios.

En este escenario actual, no resulta fácil responder preguntas sobre las relaciones de un autor o institución específica, sus temas más frecuentes de investigación, los eventos en los que se publica o indicadores de actividad de las instituciones, y resulta mucho más difícil encontrar dentro de esta red colaborativa que representan los autores patrones de asociación o bien posibles temas comunes que puedan ayudar a los investigadores a encontrarse entre sí. No obstante, existen herramientas que permiten analizar las relaciones entre investigadores y su colaboración, tales como Sci2 Tool [1] o CiteSpace [2], donde a partir de un conjunto de datos permiten analizar y visualizar redes, *hot topics*, etc. Sin embargo, estas herramientas ven limitado su potencial al no poder analizar todo el conjunto de contribuciones científicas, consecuencia de la ausencia de integración de los mencionados repositorios.

Estas dificultades tienen como causa raíz la implementación actual de la web, centrada en la presentación de información, también conocida como una web de documentos, insuficiente para la descripción de entidades contenidas y sus relaciones con otras fuentes de información [3]. Por tanto, es necesario superar algunos problemas inherentes a la naturaleza no estructurada y heterogénea de la Web y, como tal, de los repositorios mencionados.

Para superar estos obstáculos, *Linked Data* se presenta como el paradigma para la codificación, publicación e interconexión de datos estructurados entendibles para humanos y máquinas [4], de forma que la información pueda ser más fácilmente consultada en la Web Semántica. Además, este paradigma representa una nueva forma de publicar información en la Web que, hasta el momento, se ha concentrado en volcados de datos sin mayor procesamiento (XML, CSV) o en la *tradicional* publicación mediante el lenguaje HTML (*HyperText Markup Language*), con

[1] https://www.scopus.com/

[2] https://scholar.google.com.co/

[3] https://www.elsevier.com/

[4] https://www.researchgate.net/

lo que la información en la Web carece de estructura y semántica. En resumen, *Linked Data* usa la web para crear enlaces entre diversas fuentes, formando conjuntos de datos con un significado definido y con la capacidad de ser vinculado a otros conjuntos de datos existentes en la Web [5].

El potencial que brinda *Linked Data* es incuestionable, desde sus inicios se ha implementado en aplicaciones enfocadas a dominios específicos, tales como: medicina, gobierno o estadística, así como en aplicaciones web que proveen funciones orientadas al consumo de *Linked Data* (buscadores) o su manipulación [6]. Asimismo, en el ámbito temático de esta propuesta, el W3C señala la importancia de transformar los datos de repositorios de publicaciones para que estos sean "compartibles, extensibles y fáciles de reutilizar" [7]. Como consecuencia han surgido diferentes trabajos donde se mencionan los beneficios de relacionar las publicaciones científicas y *Linked Data* dentro de diferentes aplicaciones como, por ejemplo, la recuperación de datos sobre autores para la generación de su perfil [8], así como diversas iniciativas, entre las que destacan *Digital Bibliography & Library Project* (DBLP), Elsevier, etc. Sin embargo, ninguna de estas iniciativas recoge de forma exhaustiva las publicaciones científicas y las relaciones de los investigadores de la Facultad de Ingeniería de la Pontificia Universidad Javeriana.

Considerando estos aspectos, este trabajo pretende aprovechar todas las capacidades brindadas por este paradigma y explotar su potencial en la creación de una base de conocimiento científico conforme a los principios de *Linked Data*. De esta manera, en el presente trabajo se realiza un proceso de integración de las publicaciones científicas de profesores de la Universidad Javeriana, concretamente, de aquellos profesores asociados a la Facultad de Ingeniería, que están presentes en diferentes repositorios científicos. Para ello, se lleva a cabo un preprocesamiento de los datos recuperados de los mencionados repositorios, con el objetivo de realizar una limpieza de los múltiples y heterogéneos datos recopilados. Tras ello, se procede a la construcción de una red de ontologías que será utilizada para realizar una integración semántica de la información recuperada en los diferentes repositorios. Esta integración permite generar una base de conocimiento de las publicaciones científicas de los mencionados profesores de la Universidad Javeriana. Sobre esta base de conocimiento se realizan diferentes consultas para la explotación de los datos por medio de *SPARQL Protocol and RDF Query Language* (SPARQL) [9]. Estas consultas van a permitir la identificación de información sobre autores y sus publicaciones, así como de las relaciones existentes entre los mismos, y brinda la posibilidad de generar métricas y diversas visualizaciones que permiten analizar el estado de la productividad científica de la comunidad académica de la Facultad de Ingeniería de la Universidad Javeriana.

.

Este documento se estructura de la siguiente manera: En el capítulo I se recoge la descripción general del proyecto, la oportunidad del mismo y la problemática en la que se enmarca. El capítulo II contiene la descripción detallada, objetivos, metodología seguida y el potencial de innovación que tiene el proyecto. En el capítulo 3 se encuentra el marco teórico del contexto, donde se presentan los conceptos relacionados con la Web Semántica y herramientas utilizadas para el desarrollo de este trabajo. El capítulo IV está conformado por una recopilación de diferentes trabajos relacionados. En el capítulo V se describe la solución entregada con su arquitectura y las fases del desarrollo según la metodología. En el capítulo VI se presentan las validaciones adelantadas sobre el desarrollo del proyecto. Finalmente, dentro del capito VII se presentan las conclusiones, aportes y trabajo futuro concernientes al proyecto desarrollado.

I. DESCRIPCIÓN GENERAL

1.1 Oportunidad y problemática

El proyecto se centra en la recuperación de información múltiple y heterogénea sobre publicaciones científicas de autores afiliados a la Facultad de Ingeniería de la Universidad Javeriana. Por medio de consultas a diversas fuentes de información disponibles en Internet y de su correspondiente transformación a RDF, consiguiendo la integración de datos en un repositorio de publicaciones científicas que cumple con los principios de *Linked Open Data.*

Parte del desarrollo de la solución consiste en atender los problemas inherentes al contexto de las publicaciones científicas en la actualidad y a la búsqueda de información sobre publicaciones. Esta información es, generalmente, producto de consultas dentro de repositorios indexados (gratuitos y de suscripción), desconectados entre si desde los que se recuperan resultados según parámetros de entrada como el título de la publicación, el nombre del autor o el tema. La mayor parte de estas búsquedas y revisión de resultados debe hacerse de forma manual por parte del usuario, revisando diferentes fuentes y confirmando la pertinencia de los resultados por medio del cruce de referencias, verificación del contenido, revisión del nombre de los autores, identificación de resultados duplicados entre diferentes fuentes, entre otras tareas. Para las organizaciones es también problemático hacerle seguimiento a las publicaciones que generan, los temas que se están tratando y las relaciones entre investigadores.

Considerados los problemas presentes en el contexto de las publicaciones científicas se cuenta con la oportunidad de generar un repositorio de acceso libre con datos integrados y estructurados, que se alimenta de forma automática a partir de consultas a fuentes heterogéneas de Internet, con la capacidad de integrarse con otros repositorios que pueden ser de publicaciones o de cualquier otra naturaleza (geográficos, redes sociales, redes empresariales, etc.). Esta visión integrada de los diferentes repositorios le da sentido a la decisión de seguir los principios de *Linked Open Data.* Como característica adicional la integración a través del servidor SPARQL *Fuseki* permite el uso de los datos contenidos en el repositorio para aplicaciones de analítica de datos por medio de R u otras herramientas enfocadas en análisis de datos para la generación de todo tipo de métricas y visualizaciones de resultados.

II. DESCRIPCIÓN DEL PROYECTO

En esta sección se describen los objetivos del proyecto y las fases que se ejecutan para alcanzar dichos objetivos enmarcados en la metodología de publicación de *Linked Open Data.*

2.1. Objetivo general

Disponer un repositorio de conocimiento científico que brinde información de publicaciones científicas, autores y áreas de conocimiento conforme a las principales características de la Web Semántica y los principios de *Linked Data*.

2.2 Objetivos específicos

- Seleccionar y analizar las fuentes de datos relacionadas con repositorios de publicaciones.
- Desarrollar una red de ontologías que modele el dominio de conocimiento relacionado con las publicaciones científicas.
- Realizar la integración semántica de las diferentes fuentes de datos tratadas conforme a los principios de *Linked Data*.
- Desplegar una base de conocimiento de publicaciones y explotar los datos generados.

2.3 Metodología

El desarrollo de este trabajo se ciñe a la metodología de *Linked Open Data* (LOD) [10]. Para comprender lo que conlleva la mencionada metodología, se procede a contextualizar los cuatro principios básicos de *Linked Data* que se deberán observar durante este proyecto [11]:

1. Usar *Uniform Resource Identifier* (URI) como identificador exclusivo para todas las cosas.
2. Utilizar *Hypertext Transfer Protocol* (HTTP) para las URI para que sean accesibles, entrelazables e interpretables.
3. Ofrecer información sobre los recursos usando estándares abiertos como RDF (*Resource Description Framework*) y SPARQL.
4. Agregar enlaces a otros URI para su vinculación con otros datos incluidos en LOD.

Básicamente estos principios sugieren que toda URI HTTP, se pueda resolver en la Web, de modo que quien lo necesite, pueda obtener conocimiento adicional accediendo a dichos enlaces.

La metodología de *Linked Open Data* [10] adoptada en este trabajo contempla cinco actividades que van desde la descripción de los datos disponibles hasta su explotación (Figura 1). A continuación, se describen cada una de las fases desarrolladas en el proyecto, en relación con la metodología adoptada, presentando el detalle y el aporte de cada una.

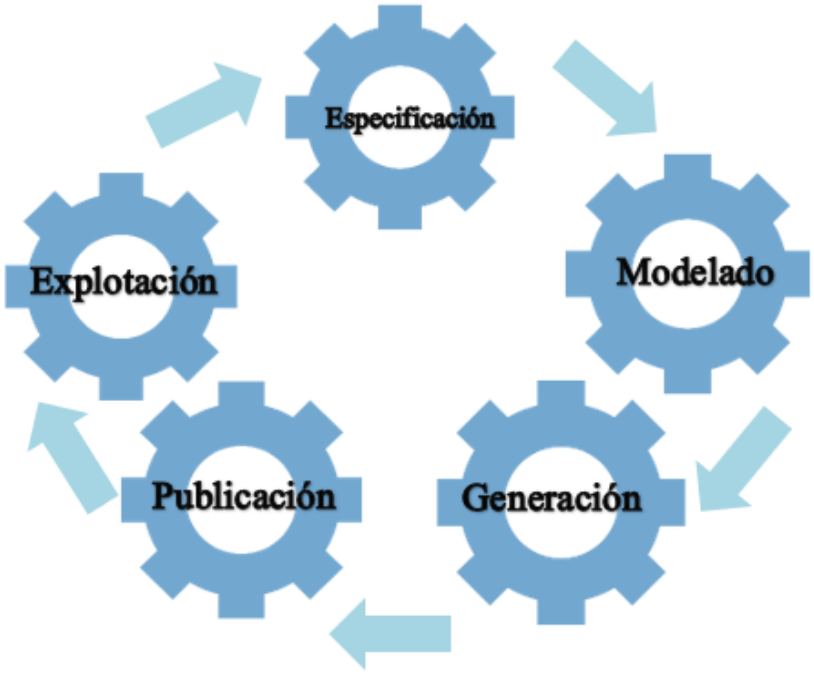


Figura 1. Fases de la metodología

2.3.1 Especificación

En esta actividad de la metodología se realiza la definición y análisis de las fuentes de datos a utilizar, diseño de las URI a utilizar y definiciones relacionadas con licenciamiento requerido. En el contexto de este trabajo se desarrollarán las siguientes tareas:

- Análisis de estructura y contenido de las API de *IEEE*, *DBLP* y *Semantic Scholar* y de dos sitios web asociados a la Facultad de Ingeniería de la Universidad Javeriana y *Google Scholar*.
- Desarrollo de una herramienta tecnológica (*crawler*) capaz de recuperar y almacenar los datos consultados.
- Limpieza de datos obtenidos de los diferentes repositorios.
- Diseño de un repositorio que pueda almacenar de forma persistente los datos resultado de la limpieza y pueda soportar los siguientes pasos de la metodología.

2.3.2 Modelado

En esta actividad se aborda el desarrollo o reutilización de vocabularios adecuados para los objetivos de un proyecto determinado. Para llevar a cabo esta tarea resulta conveniente realizar una búsqueda previa de posibles vocabularios susceptibles de ser reutilizados, de lo contrario, es decir, si no existen vocabularios reutilizables, se tendrán que construir vocabularios propios.

En el contexto de este trabajo, para el modelado se realiza la siguiente actividad:

- Ejecución de las tareas contempladas en la metodología NeOn según el escenario correspondiente [12]

2.3.3 Generación

Esta actividad conlleva un proceso de transformación de los datos originales a RDF. Para ello, se requiere utilizar el vocabulario generado o reutilizado en la fase de modelado. En el contexto de este trabajo, para la generación de RDF se realizan las siguientes tareas:

- Selección de una herramienta tecnológica que permita la transformación de los datos originales a RDF.
- Establecer las correspondencias entre los datos originales y la red de ontologías desarrollada, lo que va a permitir obtener la integración semántica de los diversos y heterogéneos datos originales recuperados.
- Generación del RDF como resultado de las transformaciones de los datos originales.

2.3.4 Publicación

La actividad de publicación conlleva el despliegue y publicación de una base de conocimiento de los datos recuperados y los metadatos que los describen, así como asegurar que estos sean descubribles desde Internet. En el contexto de este trabajo se realizan las siguientes tareas:

- Identificación de una herramienta tecnológica capaz de almacenar RDF y desplegar un SPARQL *Endpoint.*
- Publicación de los datos RDF generados sobre autores y publicaciones científicas.

2.3.5 Explotación

En la actividad de explotación se realiza un aprovechamiento de la exposición de los datos a través del servidor SPARQL (base de conocimiento). Así, en el contexto de este trabajo se realizan las siguientes tareas:

- Integración y extracción de datos para su análisis por medio de aplicaciones en la capa de presentación utilizando programas escritos en R para la generación de métricas y visualizaciones sobre diversos aspectos asociados con los autores y sus publicaciones.

2.4 Potencial de innovación

El desarrollo adelantado en este proyecto permite mejorar la forma en la que se accede al conocimiento sobre publicaciones científicas asociadas a la comunidad científica de la Facultad de Ingeniería de la Universidad Javeriana. Así, este trabajo facilita el acceso a la información sobre la producción científica por parte de docentes y directivos facilitando el proceso de toma de decisiones a través de las métricas ofrecidas, la socialización del trabajo de los investigadores y la posibilidad de encontrar patrones, redes de autores o temas comunes para las investigaciones de la comunidad Javeriana.

Otro gran aporte que brinda este trabajo se centra en los desarrollos realizados, los cuales pueden permitir generalizar estos resultados al conjunto de las Facultades de la Universidad Javeriana, permitiendo construir un repositorio propio con las publicaciones del conjunto de la comunidad, obteniendo a partir las mismas métricas y visualizaciones que faciliten el análisis de la producción científica en la Universidad. A su vez, este repositorio puede ser conectado con otras universidades siendo un potencial punto de inicio para proyectos de integración de datos sobre publicaciones científicas a nivel nacional, regional e internacional bajo el estándar de *Linked Data.*

III. MARCO TEÓRICO

En esta sección se presentan los fundamentos teóricos que dan soporte al desarrollo de este proyecto, conceptos básicos sobre las metodologías seguidas y herramientas utilizadas.

3.1 Web Semántica

La Web en su estado actual es una red de relaciones entre documentos HTML o bien un repositorio global de archivos identificados por URL (*Uniform Resource Locator*) y por dominios e IP para identificación de las máquinas, en la que tanto individuos como organizaciones comparten nuevos contenidos constantemente. En este contexto, encontrar la naturaleza de las relaciones entre ellos es cada vez más difícil debido a su rápido crecimiento y la heterogeneidad de los recursos compartidos, por esta razón los buscadores tienen problemas para encontrar resultados acertados, una gran cantidad de resultados no son pertinentes y se requiere de una revisión manual por parte del usuario, lo que tiene sentido, pues este es un modelo orientado al consumo, publicación y navegación por parte de humanos. Por esta razón surge la Web Semántica, que conlleva la construcción de una red de relaciones entre datos en la que se navega a través de sujetos, predicados y objetos, identificados por URI (*Uniform Resource Identifier*) en lugar de URL, lo que brinda una mayor información y ayuda a que las búsquedas tengan mayor precisión gracias a la estandarización de los datos [13] y su descripción por medio de RDF (*Resource Description Framework*), presentando un paradigma más apto para la navegación por parte de la máquinas y haciendo posible la representación de datos en formatos que pueden ser procesados y analizados por algoritmos gracias al lenguaje de consultas SPARQL.

Arquitectura de la Web Semántica

A continuación, se proporciona una breve descripción de los elementos que conforman la arquitectura de la Web Semántica representada en la Figura 2.

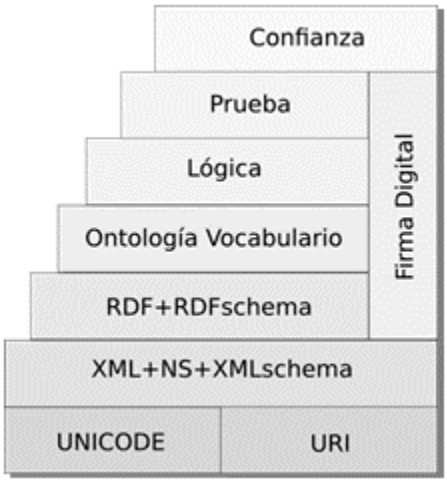


Figura 2. Arquitectura de la Web Semántica [14]

- **Unicode**: conjunto de caracteres estándar utilizado para codificar los datos.
- **URI**: identificador de recursos único utilizado para identificar cosas y conceptos.
- **XML**: lenguaje de etiquetado utilizado para el intercambio de datos.
- **RDF**: *framework* de descripción de recursos equivalente a HTML en la web semántica.
- **Ontología** – **Vocabulario**: OWL (*Web Ontology Language*), es el lenguaje utilizado para la descripción de ontologías gracias a las cuales es posible clasificar objetos y sus relaciones.
- **Lógica**: reglas de inferencia y unificación de las reglas de ontologías y significado de los datos que hacen posible el uso de razonadores lógicos.
- **Prueba**: pruebas escritas en el lenguaje unificador que se intercambian para posibilitar las inferencias lógicas.
- **Confianza**: red de confianza para las fuentes de datos y servicios, funcionamiento correcto de los sistemas, la tecnología y la interfaz de usuario.
- **Firma digital**: encriptación para la verificación de la fuente que ofrece cierta información [15].

Lenguajes Semánticos

El uso de la Web Semántica se consigue a través del uso de lenguajes capaces de describir el conocimiento por medio de metadatos y ontologías. Los principales lenguajes son RDF, RDF-S y OWL, soportados por SPARQL, el leguaje de consultas para la Web Semántica. A continuación, se proporciona una breve descripción de cada uno de estos lenguajes.

- **RDF**: Es un estándar basado en XML para el intercambio de datos RDF (*Resource Description Framework*), permite la unión y vinculación de datos, soporta cambios en el tiempo y los incorpora a la estructura del documento sin necesidad de cambiar los datos [16]. Se basa en la descripción del grafo conformado por de tripletas sujeto (recurso) predicado (propiedad) y objeto (valor) cada uno identificado por una URI.

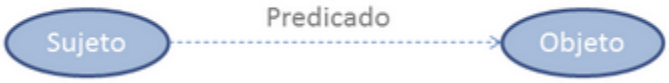


Figura 3. Tripleta RDF [17]

- **RDF-Schema**: Es una extensión de RDF que proporciona un vocabulario de modelado para datos en formato RDF, permite la representación de características de los datos y sus relaciones, como el dominio y rango o jerarquías y similitudes (*equal to, subclass of*). Se agrega en un espacio llamado RDF-S dentro del documento [18].

- OWL: Es un lenguaje para la descripción de la ontología en el que se pueden representar significados de términos, relaciones, propiedades y cardinalidad.

- SPARQL: Es un lenguaje de consulta que se utiliza para la recuperación de datos heterogéneos representados en formato RDF. Se define con el objetivo de cumplir con casos de uso específicos tales como integración y agregación de datos (consulta de varias fuentes RDF), consulta de tripletas no existentes, búsqueda a nivel de *substrings*, consultas booleanas, y retorno de resultados ordenados [19], funciona como una consulta sobre grafos en la que se encuentran similitudes y se retorna un subgrafo con los resultados.

Linked Open Data

La iniciativa *Linked Open Data* está asociada al conjunto de mejores prácticas para la publicación de datos en la Web a través de la creación de tipos de *links* entre diferentes conjuntos de datos que podrían hacer parte de diferentes organizaciones localizadas en sitios geográficos dispersos o bien entre bases de datos que requieren interoperar a nivel de datos. En general, *Linked Data* se refiere a datos publicados en internet de tal forma que pueden ser entendidos por las máquinas, por lo que tienen asociado un significado explícitamente definido, están vinculados a otros conjuntos de datos y pueden ser vinculados por otros conjuntos de datos externos [3]. A continuación, se presenta en la Figura 4 los conjuntos de datos que se encuentran publicados en la nube de *Linked Data*.

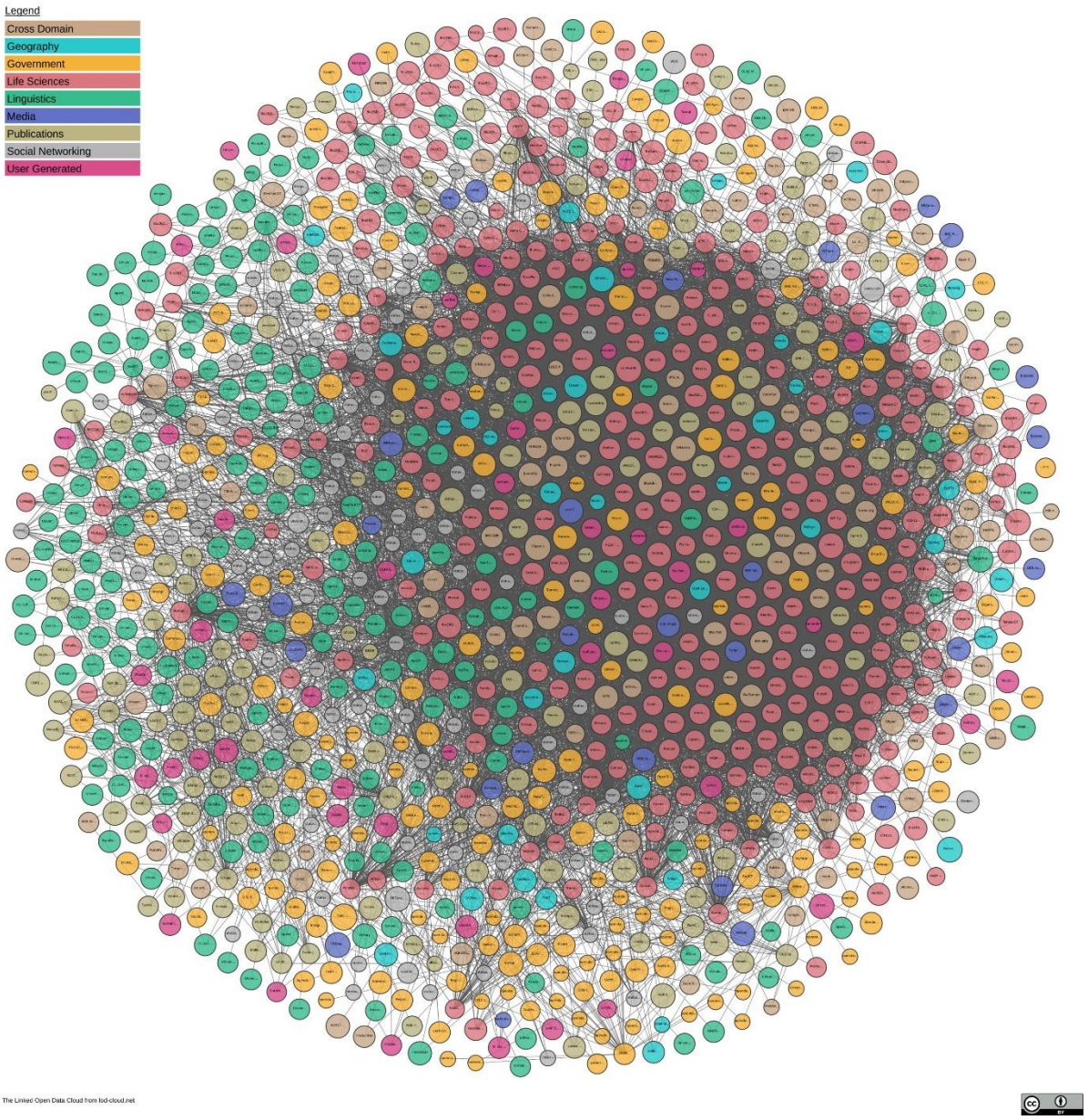


Figura 4. Linked Open Data[5].

[5] https://lod-cloud.net/

3.2 Herramientas

A continuación, se describen brevemente las herramientas tecnológicas y artefactos de software utilizados para el desarrollo de este proyecto.

ETL

Proceso de software para extracción, transformación y carga de datos comúnmente utilizado en medio de procesos de *data warehouse*, migración de datos e integración de repositorios. Pueden ser desarrollados en diferentes lenguajes de programación y herramientas de desarrollo como Talend[6], Oracle Data Integrator[7], Pentaho[8], entre otros.

Crawler

Programa que visita páginas web y recupera información de estas, pueden estar escritos en diferentes lenguajes de programación como Java y Python. Su aplicación más frecuente es como parte de motores de búsqueda.

Pentaho

Herramienta de integración de datos, orquestamiento y agendado de transformaciones (ETL) basada en JAVA. Permite acceso a fuentes homogéneas de información, en formatos variados como CSV, Excel, XML, JSON y bases de datos. Contiene diferentes pasos para la manipulación de datos preconfigurados, así como pasos contenedores de código JavaScript, cuenta con herramientas para la representación de datos e inteligencia de negocios.

LOD-GF

Plataforma para la generación de datos conforme a *Linked Open Data* que consiste en un conjunto de pasos creados para la herramienta *Pentaho* aplicadas a las diversas actividades de la metodología de Linked Data adoptada para el desarrollo de este trabajo, tales como especificación, modelado, publicación y explotación de datos.

[6] https://www.talend.com/

[7] https://www.oracle.com/middleware/technologies/data-integrator.html

[8] https://www.hitachivantara.com/go/pentaho.html

IV. TRABAJOS RELACIONADOS

Diversos autores y organizaciones han abordado el tema de los repositorios de publicaciones científicas, siendo la aproximación más común la generación de un repositorio estructurado e indexado de publicaciones que tienen algún nivel de relación que puede ser, regional, temático o institucional [20], es el caso de CARL (Colorado Alliance of Research Libraries) [21], repositorio concebido como una alianza regional para la recopilación de publicaciones científicas y soportado por la plataforma Fedora Commons[9], que cuenta con soporte para *Linked Data* y donde los metadatos son descritos utilizando XML.

Por parte de la aproximación para temas específicos se encuentran disponibles repositorios como *Collection of Computer Science Bibliographies*[10](ver Figura 5), que concentra publicaciones relacionadas con ciencias de la computación gracias al uso de herramientas como *Hot Bot*[11] y *Altavista* y la colaboración de sus usuarios, que pueden generar sus propias bibliografías y agregarlas al repositorio general. También dentro de esta categoría se encuentran los repositorios de publicaciones de la *IEEE*, *DBLP* y *Semantic Scholar*, dedicados a temas de ciencias de la computación, y en el caso de *Semantic Scholar* incluye, adicionalmente, temas médicos.

Asimismo, entre los trabajos relacionados se encuentra a EndNote [22], un producto de pago que brinda acceso un repositorio con inclinación hacia temas organizacionales. También se encuentran repositorios institucionales en los que se recuperan publicaciones según la afiliación de los autores como cVlac [23], un repositorio en el que es posible encontrar las hojas de vida de investigadores de Latinoamérica y el Caribe con información sobre sus publicaciones.

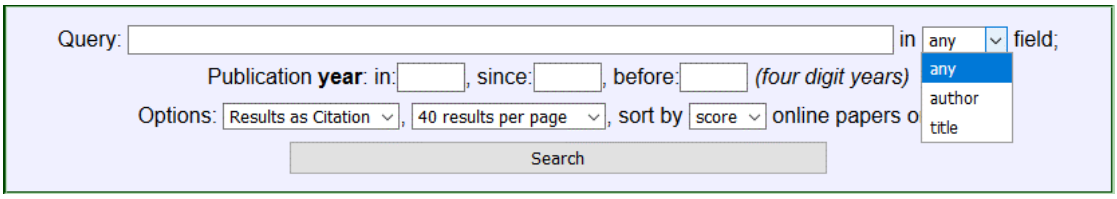


Figura 5. Búsqueda en *Collection of Computer Science Bibliographies*

Además de repositorios existen también herramientas de análisis bibliográfico, tales como, SCi2 Tool [24] y CiteSpace [25]. Ambas soluciones entregan información de co-citación, tendencias y *clusters* de autores, ayudando a encontrar investigadores interesados en temas similares a partir de conjuntos de datos específicos o bien una fuente definida.

[9] https://duraspace.org/fedora/

[10] https://liinwww.ira.uka.de/bibliography/index.html#about

[11] https://www.hotbot.com/

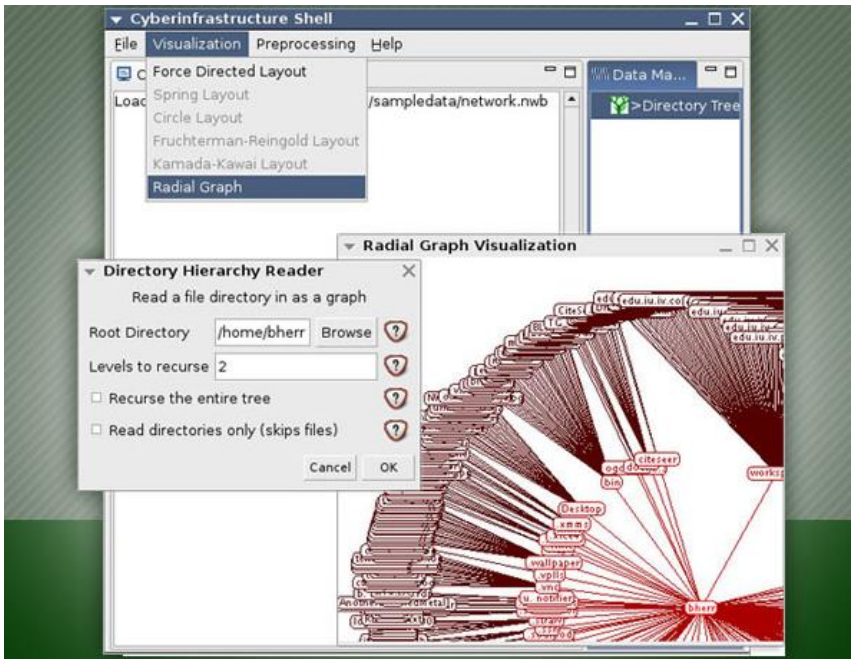


Figura 6. SCi2 Tool

El campo de la gestión documental y de las bibliotecas son uno de los focos de interés dentro la iniciativa Linked Data. Esto permitió la creación del grupo de trabajo como Linked Data Incubator Group[12] en el contexto de la W3C. Este grupo promueve el incremento y fomento de la interoperabilidad de datos bibliotecarios, la identificación de metadatos y esquemas de metadatos relacionados con bibliotecas, así como de estándares y practicas existentes para su reorientación a la Web de Linked Data. Para la descripción de lugares geográficos relacionados con recursos bibliográficos la *Library of Congress* de Estados Unidos tiene publicado su vocabulario *List for geographic areas*[13] y para el área de documentos gráficos desarrollaron el *Thesaurus of graphic materials*[14] para la descripción de imágenes.

Por otro lado, diferentes bibliotecas han desarrollado iniciativas relacionadas con *Linked* Data, como, por ejemplo, la Biblioteca Nacional de Alemania, que tiene en marcha el proyecto *Linked Data Service of the German National Library*. Este proyecto busca exponer todos los datos

[12] https://www.w3.org/2005/Incubator/lld/charter

[13] https://www.loc.gov/marc/geoareas/

[14] https://www.loc.gov/rr/print/tgm1/iib.html

bibliográficos y de autoridades integrando el conjunto de datos de bibliográficos con la Web Semántica[15].

Finalmente, otro trabajo relacionado es el buscador sueco de publicaciones Libris[16], que implementa *Linked Data* para la publicación de información sobre títulos de recursos bibliográficos pertenecientes a Universidades suecas, librerías de investigación y bibliotecas.

En general las soluciones mencionadas permiten encontrar coincidencias con los criterios de búsqueda que pueden ser título de la publicación o un tema específico y generar resultados útiles para iniciar procesos analíticos, pero no brindan acceso directo a estos datos, ni los proporcionan conforme a un estándar internacional. Cada uno tiene sus propias formas de abordar la generación de metadatos y de presentar la información, por lo que son soluciones enmarcadas en la Web de los documentos, es decir, propuestas HTML o no llegan a cumplir plenamente con los principios de *Linked Open Data*, en contraste, es posible encontrar varias formás de aplicaciones en otros contextos que vienen tomando fuerza ya que representan una alternativa robusta, relevante y con un enorme potencial de crecimiento.

[15] http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node.html

[16] (http://librishelp.libris.kb.se/

V. DESARROLLO

Dentro de esta sección se hace una descripción de la solución propuesta, su arquitectura y las diferentes actividades, herramientas y desarrollos adelantados para conseguir su implementación, así como su relación con las metodologías empleadas.

5.1 Descripción de la solución

Atendiendo las necesidades y oportunidad de mejora presente en el escenario descrito se presenta un repositorio de publicaciones científicas conforme a *Linked Open Data*. Este repositorio es alimentado de forma automática por fuentes diversas disponibles en Internet, orquestando el uso de diferentes tecnologías para la recuperación, almacenamiento y transformación de datos a formato RDF, posteriormente es expuesto a través de un servidor *SPARQL*, que permite desplegar una base de conocimiento con los datos de publicaciones integrados semánticamente. El mencionado repositorio contiene información sobre publicaciones científicas de los profesores de la Universidad Javeriana que hacen parte de la Facultad de Ingeniería, adscritos a los Departamentos de Sistemas, Industrial, Civil y Electrónica. Dicho repositorio permite acceder a información recuperada desde los datos no estructurados procedentes de diversas fuentes gratuitas en la red, siendo posible una integración con nuevas fuentes, una explotación directa por parte de diversos tipos de usuarios y un enlace con otros repositorios de conocimiento que sigan el estándar de *Linked Open Data*.

Como productos de este proyecto se crean una serie de *workflows*, desarrollados en *Pentaho*, para la transformación de datos no estructurados a datos conforme a los principios de *Linked Open Data*. Estos *workflows* pueden ser extendidos tanto en las fuentes consideradas, como en los datos que se recuperan. Además, se generan algoritmos para el preprocesado de los datos. Por otra parte, producto de la integración semántica de la información de los diversos repositorios de publicaciones se genera y publicada RDF, que puede ser utilizado en aplicaciones capaces de recuperar información a través de un SPARQL *Endpoint*. Finalmente, entre los productos desarrollados en este trabajo se encuentra una serie de *script* en R para la visualización de datos que pueden ser usados como soporte de procesos de analítica o consultas que retornan resultados pertinentes de interés para cualquier usuario interesado en la búsqueda de información sobre estas publicaciones científicas.

A continuación, se presentan las fases de la metodología *Linked Open Data* y los productos generados consecuencia del desarrollo de este proyecto (ver Tabla 1).

Tabla 1. Fases de la metodología y su producto

Fase	Producto
Especificación	Documentación de la recopilación y el análisis de datos consultados en API y sitios Web. *Workflow* de extracción de datos para cada API y de control de consultas. Algoritmo de identificación de nombres y apellidos. Algoritmo de recuperación de datos de *Google Scholar*. Algoritmo para la detección de autores javerianos.
Modelado	Red de ontologías conforme a los escenarios propuestos por la metodología NeOn.
Generación	RDF que recoge el resultado de la integración semántica de la información recopilada en los diferentes repositorios. Algoritmo para la armonización de nombres de autores. Algoritmo para la unificación de publicaciones.
Publicación	Base de conocimiento desplegada y un SPARQL *Endpoint* para su consulta.
Explotación	Visualización de datos y generación de métricas. R *Script* para la generación de métricas y visualizaciones.

5.1.1 Arquitectura

La arquitectura representada en la Figura 7 se diseña con el fin de adelantar las actividades enmarcadas dentro de la metodología de *Linked Open Data* adoptada para el desarrollo de este trabajo. Se presenta en tres componentes básicos, que son: i) extracción, entre la que se encuentran las actividades de análisis y recopilación datos; ii) persistencia, en la que se encuentra la base de datos no relacional que almacena los datos recopilados y que, posteriormente, permiten su transformación a través de un conjunto de *workflows* de Pentaho para realizar la integración

semántica de los datos recopilados; iii) presentación, en el que se integra la base de conocimiento desplegada con aplicaciones para el análisis de los datos y que la generación de visualización de datos producto del análisis.

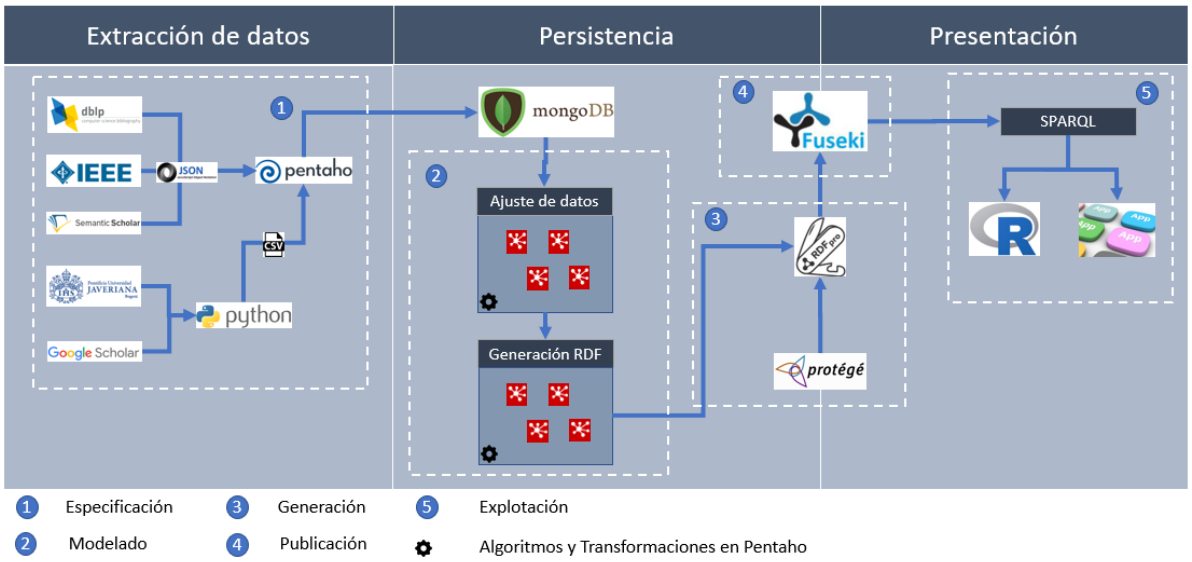


Figura 7. Arquitectura de la solución

5.1.2 Desarrollo de la solución

En esta sección se describen las fases del desarrollo contenidas en la metodología dentro de las que se explica el flujo de ejecución de los *workflows* de Pentaho, desarrollados para la consulta de datos a las fuentes, almacenamiento de resultados de las extracciones en la base de datos no relacional, las transformaciones que estos tienen para su armonización y generación de RDF. Finalmente se describen el proceso seguido para la publicación y explotación del RDF a través de ejemplos prácticos desarrollados en R.

5.1.2.1 Especificación

En esta fase se realiza un análisis de las fuentes de datos que se utilizan en el desarrollo de este proyecto. Estas fuentes se dividen en recursos Web y API, las cuales se detallan a continuación:

Recursos Web

Entre los recursos Web se encuentra la página de la Facultad de Ingeniería de la Universidad Javeriana[17]. Esta fuente tiene un rol clave en este trabajo, ya que de ella se extraen los miembros de la comunidad académica sobre los que se centra este trabajo, es decir, los autores de las publicaciones científicas. En el referido recurso Web se encuentra una sección asociada a cada uno de los departamentos que conforman dicha Facultad (Sistemas, Industrial, Civil y Electrónica). En estas secciones se presenta la información básica de los profesores adscritos a cada departamento, donde aparece detalles como: nombre, descripción, tipo de relación con la universidad y fotografía.

Para la recuperación de estos datos se desarrolla un *web crawler* en lenguaje Python, implementado a través del *framework* Scrapy [29], que proporciona herramientas para la recuperación automática de datos desde páginas web. Este *framework* permite generar como resultado un archivo de tipo CSV con los datos recopilados de los profesores de la mencionada página web.

Una vez generado el archivo de autores con su información (nombre, fotografía, relación con la universidad, descripción y departamento), se adelanta tareas de preprocesamiento de los datos. Estas tareas se enfocan, especialmente, en la información del nombre de los profesores. Las tareas realizadas incluyen la eliminación de información nula; prefijos y sufijos, tales como PhD o MSc que no hacen parte del nombre del autor; eliminación de acentos y mayúsculas, y la generación de variantes de los nombres de los autores que puedan ser usadas para su consulta en los diferentes repositorios de publicaciones científicas considerados en este trabajo. Así, se crean campos nuevos que almacenan formas alternativas de los nombres de los docentes ejecutando el Algoritmo 1 escrito en *JavaScript*:

Algoritmo 1. Identificación de nombres y apellidos

```
Entrada: Cadena de caracteres separada por espacios (autor)
Conteo de tokens que conforman la entrada:
Si el conteo de tokens es igual a 2
        Token1 = Nombre1
        Token2 = Nombre2
Si el conteo de tokens es igual a 3:
        Token1 = Nombre1
        Token2 = Apellido1
        Token3 = Apellido3
```

[17] http://ingenieria.javeriana.edu.co/profesores-

Si el conteo de tokens es igual a 4:

Token1 = Nombre1

Token2 = Apellido1

Token3 = Apellido3

Token4 = Apellido4

Si el conteo de tokens es superior a 4

Token1 = Nombre1

Token2 = Apellido1

Token3 = Apellido3

Token4 = Apellido4

Resultado: variaciones del autor para la búsqueda en las diferentes API

El resultado del Algoritmo 1 genera un objeto JSON (ver Figura 8) que será almacenado en una base de datos NoSQL, junto con toda la información recopilada.

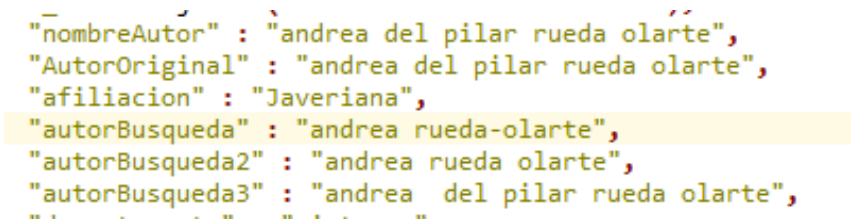


Figura 8. Variaciones del nombre autor

El segundo recurso Web tenido en cuenta como fuente de información es *Google Scholar*[18]. Esta fuente de información se considera dado que es el producto para las búsquedas de datos bibliográficos del buscador más ampliamente utilizado, siendo la primera opción en cuanto a este tipo de búsquedas. Esta fuente permite extraer información de publicaciones científicas mediante la agregación del parámetro de búsqueda "*nombre de autor*". Obteniendo como resultado los detalles de los recursos bibliográficos título, fecha, número de citaciones, URL y extracto del documento (Figura 9).

[18] https://scholar.google.com/

Figura 9. Resultado de búsqueda en Google Scholar

Para la recuperación de datos se desarrolla un algoritmo escrito en Python (ver Algoritmo 2) que permite la interacción con el módulo scholar.py [30]. Con ello se consigue la recuperación de los datos asociados a las publicaciones de los autores, que queda materializado en la generación de un archivo CSV con los resultados de las búsquedas de forma automática.

Algoritmo 2. Recuperación de datos Google Scholar

```
Entrada: Lista de autores de la Facultad de Ingeniería

Importar los módulos scholar.py y pymongo.py

Configurar conexión con la BD

Configurar query para extraer los nombres de los autores

Ciclo de ejecución en el que cada autor es consultado

        Configuración de los métodos de consulta del módulo scholar

        Ejecución de las consultas y creación de CSV con los resultados del autor

 Ciclo de consolidación de los resultados

        Lectura de cada CSV de resultados por autor

        Unión de cada CSV de autor en un CSV con resultados finales

Resultado: archivo CSV con los resultados consolidados de los autores consultados
```

Durante la ejecución de esta recuperación de datos se detecta que existe un límite de consultas diario, elemento que no aparece documentado por parte de Google en este producto. Dada esta situación, se aplica un proceso de control de consultas (ver Figura 10) desarrollado como parte del proyecto, que consiste en la ejecución del algoritmo solo para las entradas de autores que no han sido previamente consultadas. Esto se logra por medio de un *worklflow* de Pentaho que consulta los resultados obtenidos y en base a ellos actualiza la colección de entrada de autores marcándolos como consultados en la fuente *Google Scholar*. De esta manera, estos resultados

no serán tenidos en cuenta para la siguiente ejecución del Algoritmo 2. Recuperación de datos *Google Scholar*.

Adicionalmente, se detecta que los tipos de publicaciones encontradas en *Google Scholar* no corresponden exclusivamente a publicaciones científicas, ya que también se encuentra información de publicaciones de otros tipos, tales como: guías de clase, anexos de documentos de tesis, entre otros documentos. Este tipo de publicaciones también son recuperadas en este trabajo, con el objetivo dar una visión amplia de la producción de los diferentes autores. Sin embargo, estas publicaciones serán marcadas como 'Otras', para diferenciarlas de la producción científica (artículos).

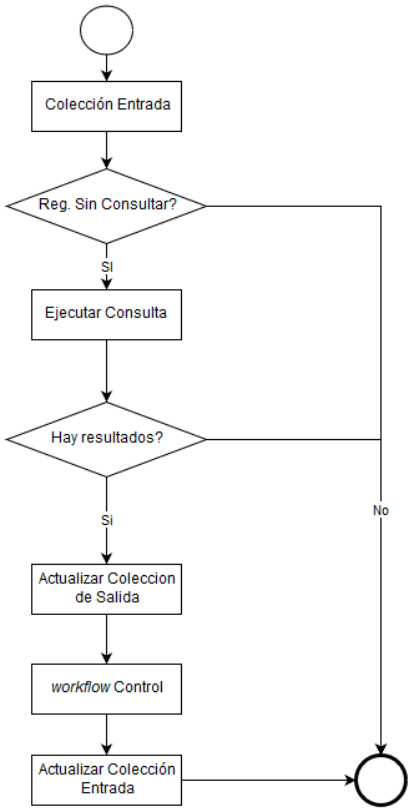


Figura 10. Generalización del proceso de consulta y control

Tras la obtención de los datos de *Google Scholar* se realiza un análisis sobre los mismos, que determina que no hay necesidad de adelantar un preprocesamiento de estos datos. Por tanto, se procede a su carga en la base de datos NoSQL desplegada en este trabajo (ver Figura 7).

Recursos sobre publicaciones

- IEEE Xplore [31]: El repositorio de publicaciones científicas de la *Institute_of_Electrical_and_Electronics_Engineers* (IEEE) presenta una fuente valiosa de información sobre publicaciones con alrededor de tres millones de documentos publicados por la IEEE y sus asociados, brinda acceso *web* y permite recuperar información a través de su API de uso gratuito, por lo que es ideal para su uso como fuente dentro de este proyecto.

  La API expuesta para la recuperación de información de sus publicaciones retorna resultados en formato JSON, cuenta con parámetros para la recuperación a partir de nombre del autor y por *Digital Object Identifier* (DOI). Bajo el primer modo de búsqueda se logra recuperar el DOI asociado a los recursos bibliográficos publicados por un autor, la segunda información detallada sobre una publicación especifica utilizando el DOI. Un ejemplo detallado de la información recuperada en esta API se muestra en el **¡Error! No se encuentra el origen de la referencia.**. API.

  Para la recuperación de los datos de esta API se desarrollan una serie de *workflows* en Pentaho, que recuperan la información asociada a los DOI de cada autor. Estos *workflows* tienen como entrada los nombres de los profesores de los diferentes departamentos de la Facultad de Ingeniería de la Universidad Javeriana, incluyendo las diferentes variaciones del nombre del autor, y una URL con los parámetros propios de la API como se observan en el siguiente ejemplo:

  ```
  http://ieeexploreapi.ieee.org/api/v1/search/articles?apikey=qu-
  mun9fy3vuq9bd6duzbdc68&format=json&max_records=25&start_re-
  cord=1&sort_order=asc&sort_field=author&author=”Nombre_del_Autor”
  ```

  Una vez obtenidos los diferentes DOI relacionados con los autores (ver Figura 11), se procede a realizar una nueva extracción de datos por medio de un nuevo *workflow en* la herramienta Pentaho. En este *workflow* la entrada será cada uno de los DOI recuperados y la URL de la API con los parámetros requeridos para su consulta. Un ejemplo de la petición realizada a esta API se muestra a continuación:

  ```
  http://ieeexploreapi.ieee.org/api/v1/search/articles?apikey=qu-
  mun9fy3vuq9bd6duzbdc68&format=json&max_records=25&start_re-
  cord=1&sort_order=asc&sort_field=author&doi=”DOI_publicación”
  ```

  Ambas consultas a la API de la IEEE se ven restringidas por la política de límites de consulta diarios de la IEEE contenida en los términos de uso[19] que puede variar según discreción de IEEE. Por ello, para estos dos procesos de extracción se aplica un control de consultas, desarrollado como parte del proyecto (Figura 10).

[19] https://developer.ieee.org/API_Terms_of_Use2

5bbd499473982b2010c47a3c	10.1109/ColumbianCC.2015.7333456	angela carrillo-ramos
5bbd499473982b2010c47a3d	10.1109/COLOMCC.2011.5936318	angela carrillo-ramos
5bbd499473982b2010c47a3e	10.1109/ICMCS.2012.6320120	angela carrillo-ramos
5bbd499473982b2010c47a3f	10.1109/ColombianCC.2012.6398030	angela carrillo-ramos
5bbd499473982b2010c47a40	10.1109/COLOMCC.2011.5936293	angela carrillo-ramos
5bbd499473982b2010c47a41	10.1109/ColumbianCC.2016.7750792	angela carrillo-ramos
5bbd499473982b2010c47a42	10.1109/ColumbianCC.2015.7333411	angela carrillo-ramos
5bbd499473982b2010c47a43	10.1109/ColumbianCC.2015.7333460	angela carrillo-ramos
5bbd499473982b2010c47a44	10.1109/COLOMCC.2011.5936315	angela carrillo-ramos
5bbd499473982b2010c47a45	10.1109/CTS.2011.5928728	angela carrillo-ramos
5bbd499473982b2010c47a46	10.1109/ColumbianCC.2016.7750780	angela carrillo-ramos
5bbd499473982b2010c47a47	10.1109/ColombianCC.2012.6398026	angela carrillo-ramos
5bbd499473982b2010c47a48	10.1109/CLEI.2012.6427212	angela carrillo-ramos
5bbd499473982b2010c47a49	10.1109/ColombianCC.2012.6398037	angela carrillo-ramos
5bbd499473982b2010c47a4a	10.1109/COLOMCC.2011.5936286	javier lopez-parra

Figura 11. Relación DOI autor

El resultado de la segunda consulta a la API de la IEEE está compuesto de los datos sobre la publicación y es almacenado sin ningún preprocesamiento dentro de la colección de publicaciones IEEE en la base de datos desplegada en este trabajo. Un ejemplo de la información almacenada se muestra en la Figura 12.


```
{
    "_id" : ObjectId("5bcbd2c163884b412c11aacf"),
    "access_type" : "LOCKED",
    "content_type" : "Conferences",
    "article_number" : "7333456",
    "doi" : "10.1109/ColumbianCC.2015.7333456",
    "title" : "Cubilletes: Adaptive system to support experiential learning in recognition and use of money",
    "publication_number" : "7317653",
    "publication_title" : "2015 10th Computing Colombian Conference (10CCC)",
    "isbn" : "New-2005_Electronic_978-1-4673-9464-2",
    "publisher" : "IEEE",
    "index_terms" : "[\"Down Syndrome\",\"adaptive system\",\"experiential \\/earning\",\"money\"]",
    "pdf_url" : "https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7333456",
    "abstract_url" : "https://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7333456",
    "html_url" : "https://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=7333456",
    "authors" : "[{\"full_name\":\"Alejandra Cubillos Carvajal\",\"affiliation\":\"Estudiante de Ingenier&#x00E
    "conference_location" : "Bogota",
    "conference_dates" : "21-25 Sept. 2015",
    "partnum" : "CFP1522V-ART",
    "start_page" : "431",
    "end_page" : "438",
    "abstract" : "This work presents Cubilletes, an adaptive system to support the experiential learning and us
    "Request2" : "http://ieeexploreapi.ieee.org/api/v1/search/articles?apikey=qumun9fy3vuq9bd6duzbdc68&format=
    "rcause" : "angela carrillo-ramos",
    "full_name0" : "Alejandra Cubillos Carvajal",
    "full_name1" : "Angela Carrillo-Ramos",
    "affiliation0" : "Estudiante de Ingenier&#x00ED;a de Sistemas, Carrera de Ingenier&#x00ED;a de Sistemas, Bo
    "affiliation1" : "Departamento de Ingenier&#x00ED;a de Sistemas, Pontificia Universidad Javeriana, Bogot&#x
}
```


Figura 12. Entrada en la colección IEEE

- API DBLP [32]: *Digital Bibliography & Library Project* (DBLP) inició como un proyecto de la Universidad de Trier (Alemania), con el tiempo ha crecido y actualmente dispone de un repositorio indexado de acceso gratuito con más de cuatro millones de publicaciones y con metadatos descritos en XML [33]. Para la consulta de información se encuentran disponibles tres API de DBLP para consultas de publicaciones, eventos

y autores, siendo la de autores la utilizada para recuperar datos en formato JSON dentro de este proyecto (ver **¡Error! No se encuentra el origen de la referencia.**).

Para conseguir la recuperación de los datos a partir de la API expuesta por DBLP se desarrolla un *workflow* de Pentaho cuyas entradas serán las variantes de los nombres de los autores para su búsqueda y la URL con sus respectivos parámetros. A continuación, se muestra un ejemplo de petición a esta API:

```
http://dblp.org/search/publ/api?q=”Nombre_Autor”&format=json
```

El resultado de la consulta a la API no pasa por ningún preprocesado y se almacena directamente en la colección de publicaciones DBLP dentro de la base de datos no relacional. Un ejemplo de la información almacenada se muestra en la Figura 13.


```
{
    "_id" : ObjectId("5bbd456e73982b2010c47889"),
    "author" : "[\"Clara Mabel Solano-Vanegas\",\"Angela Carrillo Ramos\",\"Jairo R. Montoya-Torres\"]",
    "title" : "Conceptual Framework for Agent-Based Modeling of Customer-Oriented Supply Networks.",
    "venue" : "PRO-VE",
    "pages" : "223-234",
    "year" : "2015",
    "type" : "Conference and Workshop Papers",
    "key" : "conf/ifip5-5/Solano-VanegasR15",
    "doi" : "10.1007/978-3-319-24141-8_20",
    "ee" : "https://doi.org/10.1007/978-3-319-24141-8_20",
    "urlres" : "https://dblp.org/rec/conf/ifip5-5/Solano-VanegasR15",
    "rcause" : "angela carrillo-ramos",
    "Request" : "http://dblp.org/search/publ/api?q=%22angela%20carrillo-ramos%22&format=json",
    "publisher" : "dblp"
}
```


Figura 13. Entrada en la colección DBLP

- API *Semantic Scholar* [34]: Este buscador recupera información y la indexa a partir de otros repositorios, cuenta con cuarenta millones de publicaciones principalmente relacionadas con ciencias de la computación y medicina. Además, enriquece los datos sobre publicaciones agregando información relacionada con su relevancia, encontrando enlaces entre coautores y temas relacionados [35]. Esta iniciativa tiene expuesta una API para consultas de recursos bibliográficos que permite hacer búsquedas por identificadores del recurso, DOI, S2PaperId o bien por identificador del autor ArXivId, retornando datos detallados de los recursos bibliográficos en formato JSON. En el **¡Error! No se encuentra el origen de la referencia.** se recoge una muestra de la información proporcionada por esta API.

  Debido a que esta API no cuenta con un *endpoint* que permita realizar consultas utilizando el nombre del autor, el proceso de recuperación ejecutado requiere de los siguientes pasos (ver Figura 14):

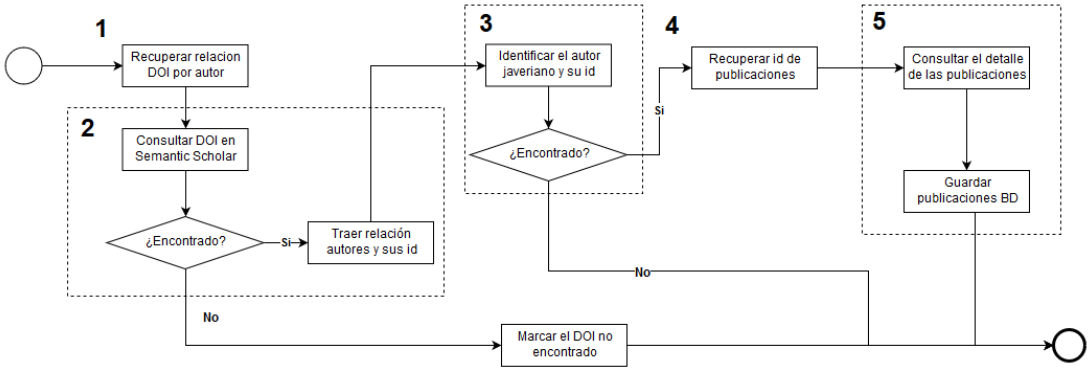


Figura 14. Flujo de recuperación *Semantic Scholar*

i) Identificación de los DOI conocidos por autor dentro de los datos de publicaciones recuperadas por otras API,

ii) Consulta de los DOI dentro de la API *Semantic Scholar* para encontrar los autores y el id de autor (ArXivId),

iii) Identificación del autor javeriano dentro de los autores recuperados por cada publicación y hallar su id de autor,

iv) Consulta con el *id* de autor, el listado de las publicaciones que tiene el autor javeriano dentro de *Semantic Scholar* y recuperar su *id* de publicación,

v) Consulta con los id de publicaciones recuperados el detalle de cada publicación.

Las consultas adelantadas para la recuperación de las publicaciones de cada autor se desarrollan usando la URL y parámetros según el ejemplo de petición (paso ii):

```
http://api.semanticscholar.org/v1/paper/”DOI_Publicacion”
```

Los resultados obtenidos para cada uno de los DOI (paso ii) son almacenados en una colección intermedia que contiene la relación de DOI, sus autores y sus respectivos *id* de autor (ver Figura 15).

```
{
    "_id" : ObjectId("5bb619bb851b19361c682f8e"),
    "rcause" : "alejandra gonzalez-correal",
    "nomAutor0" : "Miguel Angel Bermeo Ayerbe",
    "nomAutor1" : "David Stiven Avila González",
    "nomAutor2" : "Fabian Andres Merchan Jimenez",
    "nomAutor3" : "Enrique Gonzalez Guerrero",
    "nomAutor4" : "Alejandra Maria Gonzalez Correal",
    "idAutor0" : "49332547",
    "idAutor1" : "47723315",
    "idAutor2" : "33810741",
    "idAutor3" : "47618557",
    "idAutor4" : "23989465",
    "urlAutor0" : "https://www.semanticscholar.org/author/49332547",
    "urlAutor1" : "https://www.semanticscholar.org/author/47723315",
    "urlAutor2" : "https://www.semanticscholar.org/author/33810741",
    "urlAutor3" : "https://www.semanticscholar.org/author/47618557",
    "urlAutor4" : "https://www.semanticscholar.org/author/23989465"
}
```


Figura 15. Entrada Autor encontrado *Semantic Scholar*

Para identificar cuál de los *id* de autor corresponde al del autor que hace parte de la Universidad Javeriana se desarrolla el algoritmo de detección de autor javeriano escrito en *JavaScript* que tiene por entradas el campo del autor por el que se recuperó (*rcause*) y los autores que están relacionados con el DOI (paso iii). El resultado del Algoritmo 3. Detección de Autor javeriano es almacenado en una colección intermedia (ver Figura 16).

Algoritmo 3. Detección de Autor javeriano

Entrada: respuesta de la API *Semantic Scholar* consultada por DOI

Dividir campo rcause en tokens

Identificar Nombres y apellidos contenidos en rcause

Ciclo de tokenizacion y detección de nombres y apellidos dentro de los campos nomAutor

Ciclo de comparación de los nombres y apellidos de los autores Vs los nombres y apellidos del campo rcause

Salida: Nombre de autor que coincide con el autor javeriano y su correspondiente id de autor.

```
{
    "_id" : ObjectId("5bb62f72851b19361c68308c"),
    "nombreAutor" : "alejandro sierra munera",
    "rcause" : "alejandro sierra-munera",
    "semanticId" : "2377725",
    "semanticUrl" : "https://www.semanticscholar.org/author/2377725"
}
```


Figura 16. Relación entre autor javeriano e id de autor

Identificado el *id* de cada autor es posible utilizarlo para la consulta de sus publicaciones relacionadas (paso iv), por lo que se desarrolla una transformación de Pentaho que tiene por entrada la *id* de cada autor y los parámetros según el ejemplo de petición:

```
http://api.semanticscholar.org/v1/author/”Id de autor”
```

El resultado de esta consulta contiene un listado de identificadores de publicaciones que tienen relación con el *id* de cada autor, por lo que se procede a almacenarlos en una colección intermedia que pueda ser utilizada como entrada en el siguiente paso (ver Figura 17). Esta colección actuará como colección de entrada para el proceso de consulta y control (ver Figura 10).

rcause	request	paperId	estado
alejandra gonzalez-correal	http://api.semanticscholar.org/v1/author/23989465	69554e0f5fd2d9c3d051aacdfc953537db4eff68	ok
alejandro sierra-munera	http://api.semanticscholar.org/v1/author/2377725	545a84acd3ebc05ab73b1c0d486bb97d9399309c	ok
alejandro sierra-munera	http://api.semanticscholar.org/v1/author/2377725	6ac905eb9c13013678b5f7d6d73cd7055b417e17	ok
alejandro sierra-munera	http://api.semanticscholar.org/v1/author/2377725	1c720594245201 72db15841e4fc844d8da268e8c	ok
alejandro sierra-munera	http://api.semanticscholar.org/v1/author/2377725	ecdccc41969c893e3480f1e0c4f1d93f96739f26	ok
alejandro sierra-munera	http://api.semanticscholar.org/v1/author/2377725	d13ace9648f7b9a8ef29312c5a1091d16cb6491c	ok
alejandro sierra-munera	http://api.semanticscholar.org/v1/author/2377725	de6b4d7ec20e9fafdf97b31f55f7856889a0939d	ok
alejandro sierra-munera	http://api.semanticscholar.org/v1/author/2377725	2859ab234362985068afdda949dda3e28e7ec22a	ok
alejandro sierra-munera	http://api.semanticscholar.org/v1/author/2377725	7056917fbd794d5fc7d21f1cb3645e6c6d237cf6	ok
alejandro sierra-munera	http://api.semanticscholar.org/v1/author/2377725	1cd9d6d1a8545f371d81f1c77cae578ff1146b5b	ok

Figura 17. Relación Autor paper *id* Semantic Scholar

En el último paso (paso v) se desarrolla una transformación de Pentaho que recupere la información detallada de cada una de las publicaciones, usando como entrada el *id* de la publicación (*paperId*) y los parámetros según el ejemplo de consulta:

```
http://api.semanticscholar.org/v1/paper/”paper_id”
```

Finalizados los pasos previamente descritos, se cuenta con datos sobre cada una de las publicaciones, los cuales son cargados dentro de la colección de publicaciones de *Semantic Scholar* en la base de datos desplegada en este trabajo. En la Figura 18 se muestra un ejemplo de la entrada de la colección recogida en la base de datos.

```
{
    "_id" : ObjectId("5bd06a61d88f5b1c94287968"),
    "rcause" : "alejandro sierra-munera",
    "publisher" : "semScholar",
    "request2" : "http://api.semanticscholar.org/v1/paper/545a84acd3ebc05ab73b1c0d486bb97d9399309c",
    "title" : "Overtraining modifies spatial memory susceptibility to corticosterone administration.",
    "doi" : "10.1016/j.nlm.2017.10.003",
    "topic1" : "Corticosterone",
    "topic2" : "Memory Disorders",
    "topic3" : "Extinction, Psychological",
    "topic4" : "PersonNameUse - assigned",
    "topic5" : "Twenty Four",
    "topic6" : "mg/kg",
    "autor0" : "Alejandro Sierra Múnera",
    "autor1" : "Mayerli A Prado-Rivera",
    "autor2" : "D Carolina Cárdenas-Poveda",
    "autor3" : "Marisol Rodríguez Lamprea",
    "venue" : "Neurobiology of learning and memory",
    "year" : "2017",
    "pdf_url" : "https://www.semanticscholar.org/paper/545a84acd3ebc05ab73b1c0d486bb97d9399309c",
    "pClave" : "Corticosterone, Memory Disorders, Extinction, Psychological, PersonNameUse - assigned, T
}
```


Figura 18. Entrada de la colección de *Semantic Scholar*

Una vez se cuenta con colecciones de publicaciones para cada una de las API, se analizan los campos recogidos en cada una y se determina los campos que tendrá la colección final producto de esta fase del desarrollo. Asimismo, estos campos se usan como entrada de las siguientes fases de desarrollo, los cuales contienen campos armonizados descritos en la Tabla 2:

Tabla 2. Campos de la colección Publicación

Campo	Descripción
publisher	Organización que publica los datos (IEEE, DBLP, *Semantic Scholar*, Google Scholar)
title	Título de la publicación
abstract	Resumen de la publicación
abstract_url	URL en la que se puede consultar el resumen
type	Tipo de publicación libro, conferencia, revista, etc.
doi	Identificador digital de objetos
isbn	Numero internacional estándar de libro
urlPublicacion	URL en la que se puede consultar la publicación
conference_location	Localización geográfica en la que se presentó la publicación
date	Fecha del evento en el que se presentó la publicación
topic[0-9]	Palabra clave
rcause	Autor javeriano por el que se recuperó la publicación
Autor[0-9]	Autor de la publicación
year	Año de la publicación
venue	Evento en el que se presentó la publicación
volume	Volumen en el que se encuentra recopilada la publicación
pages	Paginas en las que se encuentra recopilada la publicación

5.1.2.2. Modelado

En esta fase se crea una red de ontologías creada como un módulo a partir de otras ontologías existentes, para describir el dominio de conocimiento relacionado con las publicaciones científicas siguiendo las pautas marcadas en la metodología NeOn [12]. Esta metodología se basa en escenarios para la construcción de redes de ontologías en entornos distribuidos y está compuesta por un conjunto de nueve escenarios, un glosario de procesos involucrados en el desarrollo de ontologías y unas directrices metodológicas.

Los escenarios propuestos por la metodología NeOn son los siguientes:

1. Desde la especificación de la aplicación a la implementación: la ontología es desarrollada desde cero.
2. Reúso y reingeniería de recursos no ontológicos: reutilización de recursos no ontológicos que se determinen según los requerimientos y reingeniería para transformarlos a recursos ontológicos.
3. Reutilización de recursos ontológicos: uso de recursos ontológicos como un todo o como módulos.
4. Reúso y reingeniería de recursos ontológicos: los desarrolladores reúsan y reorganizan los recursos ontológicos.
5. Reutilización y unión de recursos ontológicos: cuando varios recursos ontológicos que hacen parte del mismo dominio son seleccionados para su uso y creación de nuevos recursos ontológicos a partir de la unión.
6. Reúso, unión y reingeniería de recursos ontológicos: los desarrolladores, reúsan, unen y hacen reingeniería de recursos ontológicos reorganizándolos según sus necesidades.
7. Reutilización de los patrones de diseño de ontologías: los desarrolladores acceden a repositorios de reutilización de patrones de diseño de ontologías.
8. Reestructuración de recursos ontológicos: los desarrolladores reestructuran los recursos ontológicos que deben integrarse por medio de modularizarían, poda, extensión yo especialización.
9. Localización de recursos ontológicos: adaptación de la ontología a otros lenguajes o culturas generando una ontología multilenguaje.

Considerando el contexto de este trabajo, como primer paso se detecta que el escenario más adecuado para el desarrollo de la red de ontologías es el seis "*reusar, unir y reingeniería de recursos",* dado que existen varios recursos ontológicos para el dominio de los datos recuperados enmarcado en los recursos bibliográficos y descripción de personas. Dichos recursos se sobreponen entre si y pueden ser utilizados para la construcción de una red de ontologías que brinde descripción de todos los datos considerados. Adicionalmente, se cumple con los prerrequisitos necesarios para el desarrollo de este escenario que son: contar con conocimiento del dominio de la ontología y la existencia de recursos ontológicos para el desarrollo de la red de ontologías.

Por tanto, la primera actividad de esta fase consiste en la búsqueda de vocabularios existentes con el fin de reutilizar recursos ontológicos para describir los datos. De esta manera, se procede

a la búsqueda de recursos ontológicos en el repositorio de vocabularios *Linked Open Vocabularies*[20], en donde se realizan búsquedas de términos genéricos, que de antemano se sabe son necesarios para describir los datos. De esta forma se encuentran recursos y las ontologías que los contienen (ver Figura 19).

Figura 19. Búsqueda de ontologías relevantes *Dublin Core*

Tras la realización de un proceso de búsquedas en profundidad, se encontraron tres ontologías disponibles que se utilizan por su pertinencia en relación con el contexto de este proyecto. Las ontologías seleccionadas son:

- **vCard**: Una ontología para la descripción de personas y organizaciones desarrollada por W3C [36].
- **Bibliographic Ontology**: Esta ontología modela conceptos y propiedades de citas y recursos bibliográficos [37].
- **Dublin Core**: Esta ontología describe diferentes tipos de recursos que pueden ser digitales o físicos [38].
- **FOAF**: Una ontología desarrollada especialmente para la descripción de personas, grupos y documentos [39].

Asimismo, tras un proceso de análisis de las mencionadas ontologías, se detecta que estas contienen términos de otras ontologías que han sido reutilizadas. Por tanto, también serán tenidas en cuenta para la conformación de la red de ontologías de este trabajo.

[20] https://lov.linkeddata.es

Para la manipulación y generación de la red de ontologías se utiliza la herramienta *Protégé*[21] con la que, siguiendo el escenario adoptado de la metodología NeOn, se desarrollan las siguientes actividades:

1. Poda: Esta tarea permite identificar y clasificar los recursos de cada ontología (según su pertinencia) para generar una versión más adecuada conforme a los objetivos de este trabajo (ver Figura 20).

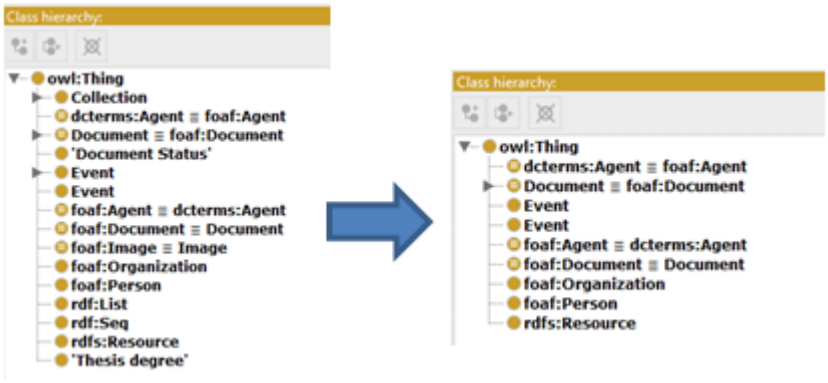


Figura 20. Poda Bibliographic

2. Unión: Para la creación de la red de ontologías, como paso previo, se realiza la importación de las diferentes ontologías a reutilizar. Esta tarea se lleva a cabo tras la realización de la poda sobre las versiones de las ontologías iniciales (ver Figura 21).

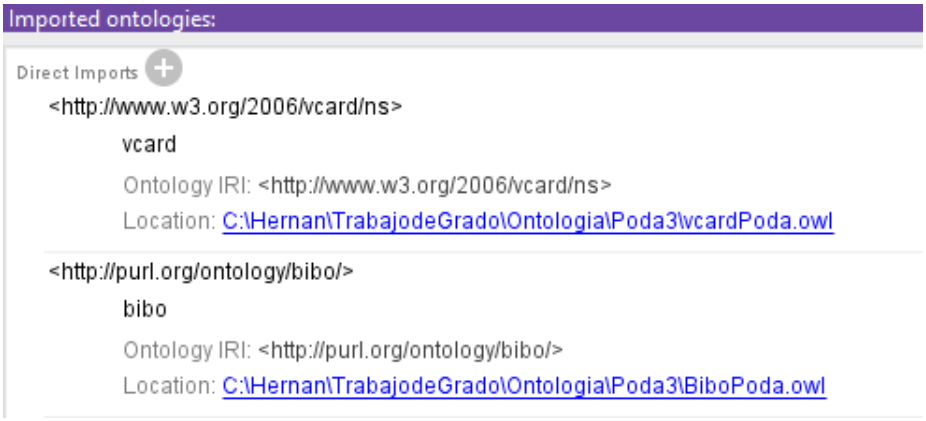


Figura 21. Importación de ontologías

[21] https://protege.stanford.edu/

3. Reingeniería (mapeo): En esta tarea se lleva a cabo una revisión y modificación del modelado del conocimiento de los diferentes recursos ontológicos considerados. Así, se realiza una inclusión de definición de axiomas, relaciones, dominio y rango, entrelazado de términos de las tres ontologías como se observa en la Figura 22 y Figura 23.

Figura 22. Entrelazado de ontologías VCard y FOAF

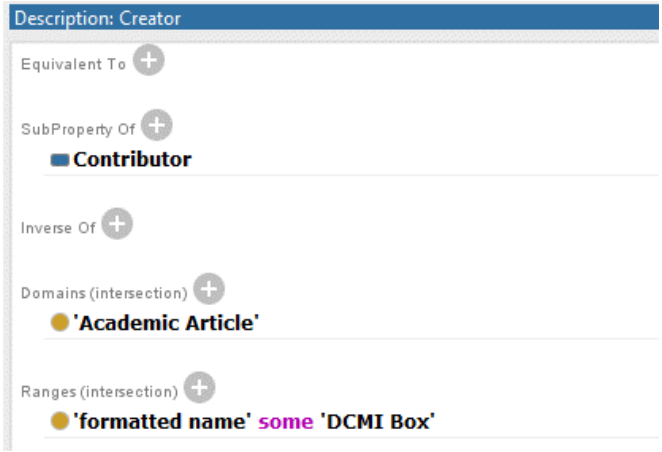


Figura 23. Definición de axioma.

La Figura 24 ilustra una visión de alto nivel de la red de ontologías creada como resultado de las tareas realizadas para el modelado del conocimiento relacionado con autores y sus publicaciones científicas.

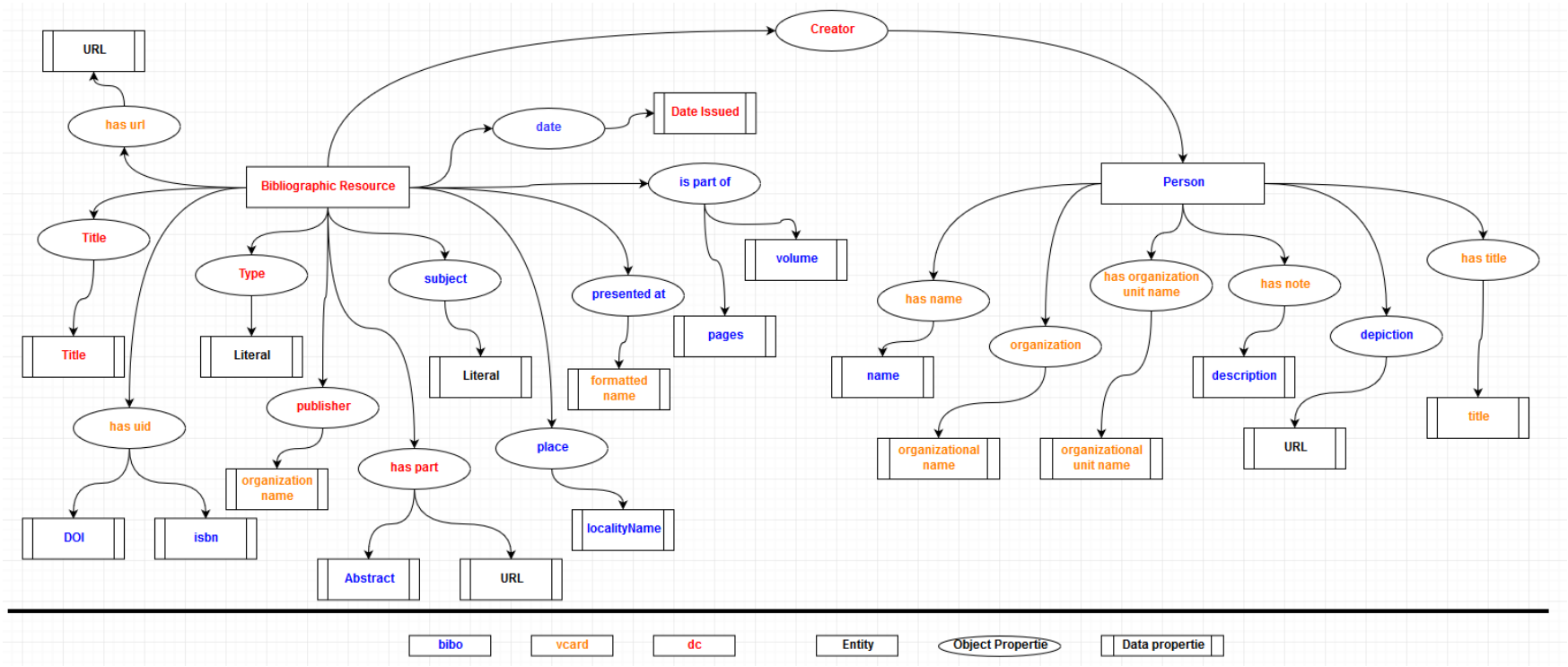


Figura 24. Modelo de alto nivel de la red de ontología desarrollada

5.1.2.3. Generación de RDF

A continuación, conforme a la metodología de *Linked Data* adoptada en este trabajo, se describen las tareas realizadas para la transformación de los datos originales, utilizando la red de ontologías desarrollada, lo que permite obtener una integración semántica de los datos obtenidos de los diferentes repositorios y generar información en RDF.

Limpieza y preprocesamiento de datos

Las publicaciones recuperadas de las diferentes fuentes tienen estructuras heterogéneas, por esta razón es necesario adelantar transformaciones para armonizar los datos y llegar a una versión final de los mismos compuesta por la estructura de una publicación definida (ver Tabla 3. Armonización de datos IEEE), por lo que para cada fuente se adelantan procesos específicos (*workflows*) desarrollados en *Pentaho* y que se describen a continuación:

- IEEE: Sobre los datos recuperados del repositorio de IEEE se realiza el siguiente proceso de limpieza: Eliminación de publicaciones sin DOI, limpieza de caracteres especiales ([,],",/) y división de palabras clave contenidas en un solo campo (*index_terms*), siendo el resultado los campos *Topic* [0-9].

  En cuanto al preprocesamiento de los datos de esta fuente, se realizan las siguientes tareas: se procede a la generación del campo *Pages*, que recoge el número de páginas de un artículo, a partir de los *start_page* y *end_page*, lo que permite la armonización de los datos. Además, se realiza un manejo de campos nulos y de los campos que se encuentran en la colección Publicación y no en la colección de la fuente de origen (IEEE), a los que se les da valor '*NA*' para evitar inconsistencias en fases posteriores. Los cambios realizados se observan en la Tabla 3 en la columna Acción, donde los campos que no aparecen fueron descartados.

Tabla 3. Armonización de datos IEEE

Publicación	IEEE
doi	doi
publisher	publisher
title	title
abstract	abstract
abstract_url	abstract_url
type	content_type
isbn	isbn
urlPublicacion	pdf_url
conference_location	conference_location

date	conference_dates
topic[0,9]	index_terms
rcause	rcause
autor[0,9]	full_name
venue	publication_title
volume	NA
pages	start_page
	end_page

- DBLP: Se realiza la eliminación de publicaciones sin DOI. Limpieza de caracteres especiales para el campo *authors* y separación, puesto que en este campo originalmente se encuentran todos los autores. Renombrado de campos para manejar un nombre común entre las diferentes fuentes. También se tratan los campos nulos y aquellos campos que se encuentran en la colección Publicación y no en la colección de la fuente de origen (DBLP) a los que se les da valor '*NA*' para evitar inconsistencias en fases posteriores. Los cambios adelantados se observan en la Tabla 4 en la columna Acción, donde los campos que no aparecen fueron descartados.

Tabla 4. Armonización de datos DBLP

Publicación	**DBLP**
doi	doi
publisher	publisher
title	title
abstract	NA
abstract_url	NA
type	type
isbn	NA
urlPublica-cion	ee
confe-rence_loca-tion	NA
date	year
topic[0,9]	NA
rcause	rcause

autor[0,9]	author
venue	venue
volume	volume
pages	pages

- *Semantic Scholar*: Se lleva a cabo una eliminación de publicación sin DOI y manejo de campos nulos y de los campos que se encuentran en la colección Publicación y no en la colección de la fuente de origen (*Semantic Scholar*) a los que se les da valor '*NA*' para evitar inconsistencias en fases posteriores. Renombrado de campos para manejar un nombre común entre las diferentes fuentes. Los cambios adelantados se observan en la Tabla 5 en la columna Acción, donde los campos que no aparecen fueron descartados.

Tabla 5. Armonización de datos *Semantic Scholar*

Publicación	**SemanticScholar**
doi	doi
publisher	publisher
title	title
abstract	NA
abstract_url	NA
type	NA
isbn	NA
urlPublicacion	pdf_url
conference_location	NA
date	year
topic[0,9]	topic1
rcause	rcause
autor[0,9]	authors
venue	NA
volume	NA
pages	NA

Los datos resultantes del proceso de limpieza y armonización son cargados a la colección *Publicaciones* en la que se adelantan tareas adicionales previas a la generación del RDF.

Eliminación de duplicados

Dado que la recuperación de publicaciones disponibles en los diferentes repositorios considerados se ejecuta en procesos paralelos y se carga en la colección *Publicaciones*, se detectan diversos problemas en los datos que se abordan de la siguiente manera:

Desambiguación del nombre del autor. Con frecuencia un artículo está conformado por diversos autores que, en ocasiones, presentan diferentes afiliaciones. Por ello, se realiza un proceso para encontrar entre los autores de cada publicación qué autor pertenece a la Facultad de Ingeniería de la Universidad Javeriana, motivo por el cual fue recuperada una publicación (ver Figura 25). Así que se realiza una desambiguación del nombre de los autores que consiste en la ejecución del Algoritmo 4. Cambio del nombre del autor javeriano, que actualiza el nombre del autor como se recupera desde la publicación cambiándolo por el nombre que se encuentra dentro la página de la Universidad Javeriana.

Este algoritmo es una solución basada en grafos dado que se conoce con antelación cual es la estructura de los nodos que representan en un primer nivel, nombres y apellidos, de los cuales se desprenden primer y segundo nombre y primer y segundo apellido, según el caso. No obstante, existen otros métodos más robustos para realizar la desambiguación de los autores como la asignación de pesos y revisión de variables correlacionadas con uno u otro autor. Sin embargo, para el alcance del proyecto resulta suficiente con la aproximación de grafos [40].

Algoritmo 4. Cambio del nombre del autor javeriano

Entrada: *campos de autores de la publicación y campo del nombre del autor javeriano por el que se recuperó la publicación (recuperado por).*

Recuperación de los campos de autores y del campo "recuperado por"

Tokenizacion del campo "recuperado por" cada token es un autor

Guardar en un arreglo los tokens

Guardar en un arreglo todos los autores de la publicación

Ciclo de comparación

> *Iterar comparando cada uno de los elementos de los dos arreglos generados*

Si existen coincidencias actualizar el campo que corresponda

Si no existen coincidencias marcar la publicación para ser descartada (se recuperó, pero ningún autor coincide con la búsqueda)

Salida: *Campos de autor de la publicación actualizados con el nombre del autor javeriano como aparece en la página de la Universidad.*

Como resultado de la ejecución el algoritmo para todas las entradas se genera una colección intermedia con los datos ajustados (ver Figura 26).


```
{
    "_id" : ObjectId("5bb75ddc09eff13c8cc9768a"),
    "publisher" : "dblp",
    "title" : "AIO robot - A EDI modular robotic dramatization platform.",
    "abstract" : "NA",
    "abstract_url" : "NA",
    "type" : "Conference and Workshop Papers",
    "doi" : "10.1109/ICAR.2017.8023528",
    "isbn" : "NA",
    "urlPublicacion" : "https://doi.org/10.1109/ICAR.2017.8023528",
    "conference_location" : "NA",
    "conference_dates" : "NA",
    "topic1" : "NA",
    "topic2" : "NA",
    "topic3" : "NA",
    "topic4" : "NA",
    "topic5" : "NA",
    "topic6" : "NA",
    "topic7" : "NA",
    "topic8" : "NA",
    "rcause" : "alejandra gonzalez-correal",
    "autor0" : "Miguel Angel Bermeo Ayerbe",
    "autor1" : "David Stiven Avila González",
    "autor2" : "Fabian Andres Merchan Jimenez",
    "autor3" : "Enrique Gonzalez Guerrero",
    "autor4" : "Alejandra Maria Gonzalez Correal",
    "autor5" : "NA",
    "autor6" : "NA",
    "autor7" : "NA",
    "autor8" : "NA",
    "autor9" : "NA",
    "year" : "2017",
    "venue" : "ICAR",
    "volume" : "NA",
    "pages" : "262-268"
}
```


Figura 25. Publicación antes de identificar al autor javeriano

```
{
    "_id" : ObjectId("5bba460eb2bd470dac8045c3"),
    "publisher" : "dblp",
    "title" : "AIO robot - A EDI modular robotic dramatization platform.",
    "abstract" : "NA",
    "abstract_url" : "NA",
    "type" : "Conference and Workshop Papers",
    "doi" : "10.1109/ICAR.2017.8023528",
    "isbn" : "NA",
    "urlPublicacion" : "https://doi.org/10.1109/ICAR.2017.8023528",
    "conference_location" : "NA",
    "conference_dates" : "NA",
    "topic1" : "NA",
    "topic2" : "NA",
    "topic3" : "NA",
    "topic4" : "NA",
    "topic5" : "NA",
    "topic6" : "NA",
    "topic7" : "NA",
    "topic8" : "NA",
    "rcause" : "alejandra gonzalez-correal",
    "autor0" : "miguel angel bermeo ayerbe",
    "autor1" : "david stiven avila gonzalez",
    "autor2" : "fabian andres merchan jimenez",
    "autor3" : "enrique gonzalez guerrero",
    "autor4" : "alejandra maria gonzalez correal",
    "autor5" : "na",
    "autor6" : "na",
    "autor7" : "na",
    "autor8" : "na",
    "autor9" : "na",
    "year" : "2017",
    "venue" : "ICAR",
    "volume" : "NA",
    "pages" : "262-268",
    "autorJav" : "4",
    "nombreAutor" : "alejandra maria gonzalez correal"
}
```


Figura 26. Publicación después de identificar autor javeriano

- Publicaciones repetidas. En ocasiones, los repositorios considerados presentan información duplicada sobre una publicación o los miembros de la comunidad científica de la Facultad de Ingeniería realizan colaboraciones entre ellos. Esto conlleva que se produzca una recuperación duplicada de la información de las publicaciones, ya que existen casos en los que la misma publicación es recuperada más de una vez producto de la consulta de uno o más autores de la Javeriana a los repositorios de referencia. Para eliminar esta duplicidad se ejecuta el Algoritmo 5 de unificación de publicaciones a través de un *workflow en Pentaho.* Este algoritmo se encarga de agrupar por el campo

DOI y selecciona los campos presentes o con un mayor contenido para generar a partir de varias entradas de la misma publicación, una única entrada que contenga la mayor cantidad de campos posible sobre una determinada publicación.

Algoritmo 5. Unificación de publicaciones

Entrada: *colección de publicaciones en la que existen duplicados con campos heterogéneos recuperados de diferentes fuentes*

Agrupar todas las publicaciones por campo DOI y los demás campos concatenarlos usando como separador ‘,’.

Tokenizacion de cada campo y comparación entre tokens

Si existe más de un token con información seleccionar el que tiene una mayor longitud y mantenerlo como valor del campo, si son iguales mantener el primero de la comparación.

Si solo un token que tiene información mantener este como valor del campo

Salida: *colección de publicaciones en la que los duplicados se unen en un registro único con la mayor cantidad de información por campos posible*

Terminadas las tareas de preparación de datos se procede al almacenamiento de esta nueva versión depurada de los datos de publicaciones en la base de datos NoSQL. Esta carga de datos recoge una colección que representa a todas las publicaciones recuperadas que será utilizada en el proceso de integración semántica, plasmado a través de la generación de RDF.

Transformación a RDF

Para la transformación de los datos obtenidos de los diferentes repositorios a RDF se utilizan como entradas las colecciones *Autores* y *Publicaciones* almacenadas en la base de datos desplegada en este trabajo. A estas entradas se le suma la red de ontologías desarrollada en la fase de modelado. A continuación, se describen los pasos seguidos para conseguir la transformación a RDF que va a permitir la integración semántica de los datos de autores y publicaciones científicas.

1. Selección de la herramienta LOD-GF[22] para la generación de RDF a partir de las colecciones. Esta herramienta se selecciona debido a que brinda las funcionalidades de

[22] https://github.com/ucuenca/lodplatform

transformación de datos a RDF y configuración de servidor *Fuseki* necesarias para el desarrollo de este proyecto. Adicionalmente, está basado en *Pentaho* que es la principal herramienta utilizada para todas las fases del desarrollo, por lo que se ajusta a las necesidades y al desarrollo adelantado.

2. Desarrollo de un *workflow* en Pentaho para llevar a cabo la transformación. En este *workflow* se utilizan diferentes *plugins* de la herramienta LOD-GF (ver Figura 27). Estos *plugins* permiten establecer las correspondencias entre los diversos conjuntos de datos y los elementos (clases, relaciones y atributos) de la red de ontologías desarrollada (ver Figura 28).

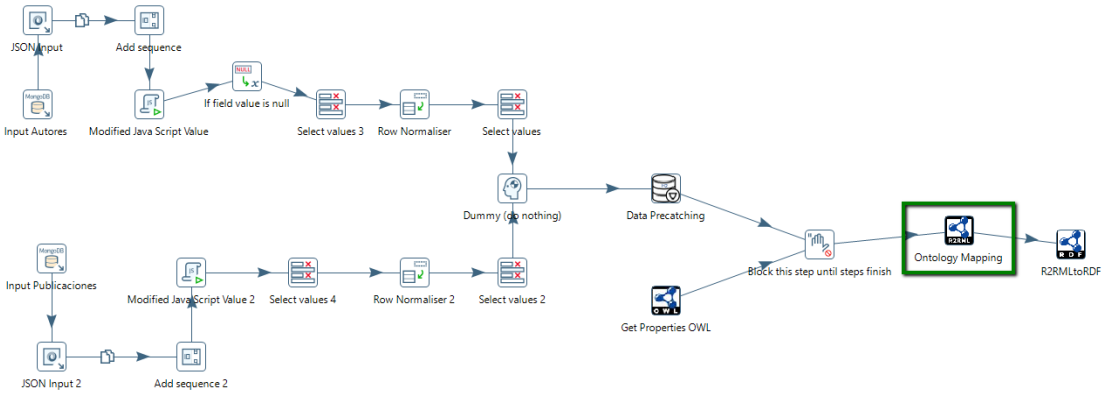
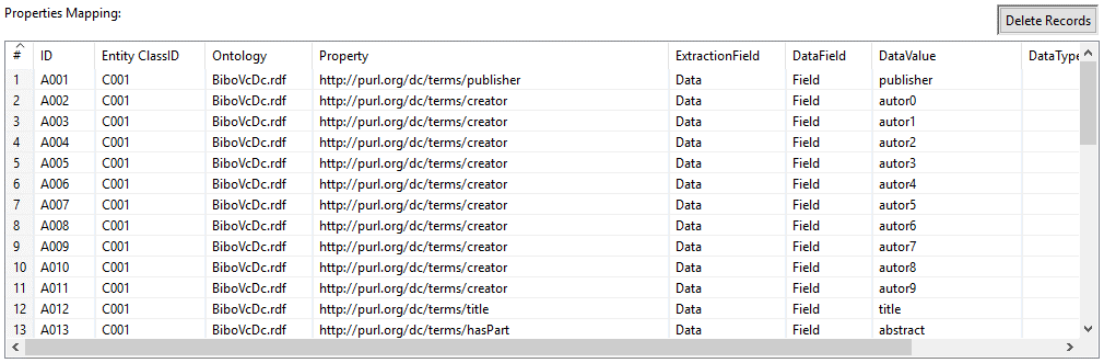


Figura 27. Transformación de JSON a RDF


Properties Mapping:

Delete Records

#	ID	Entity ClassID	Ontology	Property	ExtractionField	DataField	DataValue	DataType
1	A001	C001	BiboVcDc.rdf	http://purl.org/dc/terms/publisher	Data	Field	publisher	
2	A002	C001	BiboVcDc.rdf	http://purl.org/dc/terms/creator	Data	Field	autor0	
3	A003	C001	BiboVcDc.rdf	http://purl.org/dc/terms/creator	Data	Field	autor1	
4	A004	C001	BiboVcDc.rdf	http://purl.org/dc/terms/creator	Data	Field	autor2	
5	A005	C001	BiboVcDc.rdf	http://purl.org/dc/terms/creator	Data	Field	autor3	
6	A006	C001	BiboVcDc.rdf	http://purl.org/dc/terms/creator	Data	Field	autor4	
7	A007	C001	BiboVcDc.rdf	http://purl.org/dc/terms/creator	Data	Field	autor5	
8	A008	C001	BiboVcDc.rdf	http://purl.org/dc/terms/creator	Data	Field	autor6	
9	A009	C001	BiboVcDc.rdf	http://purl.org/dc/terms/creator	Data	Field	autor7	
10	A010	C001	BiboVcDc.rdf	http://purl.org/dc/terms/creator	Data	Field	autor8	
11	A011	C001	BiboVcDc.rdf	http://purl.org/dc/terms/creator	Data	Field	autor9	
12	A012	C001	BiboVcDc.rdf	http://purl.org/dc/terms/title	Data	Field	title	
13	A013	C001	BiboVcDc.rdf	http://purl.org/dc/terms/hasPart	Data	Field	abstract	

Figura 28. Mapeo de la ontología

3. Ejecución del *workflow* utilizando los *plugins* LOD-GF. Esta ejecución permite que los datos pasen por una serie de transformaciones dentro del flujo de ejecución de *Pentaho.* Estas transformaciones se llevan a cabo dentro de la base de datos relacional H2 embebida en JAVA, logrando una normalización de los datos que los deja preparados

para ejecutar el mapeo de recursos ontológicos y datos, que posteriormente, retorna el archivo RDF con los datos sobre autores y publicaciones científicas.

4. Unificación de RDF. Tras la obtención del RDF, se procede a la unión del RDF generado con la red de ontologías desarrollada en la fase de modelado. Para ello se utiliza la herramienta RDF-PRO[23], que mediante la función *merge*, permite unificar modelo y datos en un mismo archivo para su publicación y explotación.

5.1.2.4. Publicación y explotación

Para la publicación del RDF generado y la red de ontologías se requiere desplegar un *triple store*, es decir, una base de datos para el almacenamiento y retorno de tripletas en las que cada entrada está compuesta por sujeto-predicado-objeto del tipo "José-es un-profesor titular".

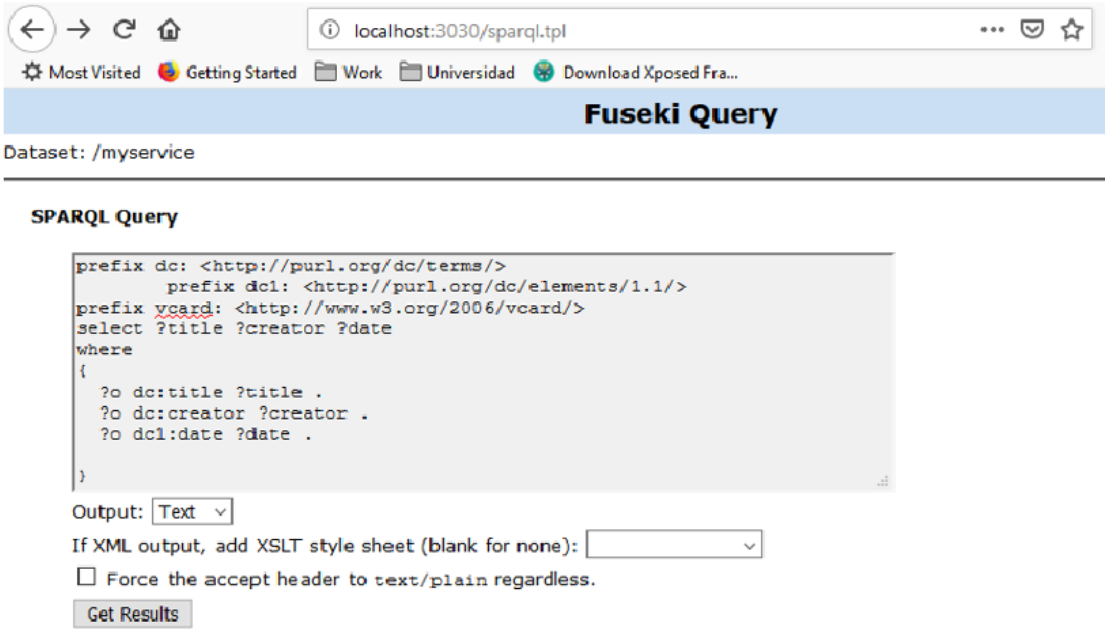
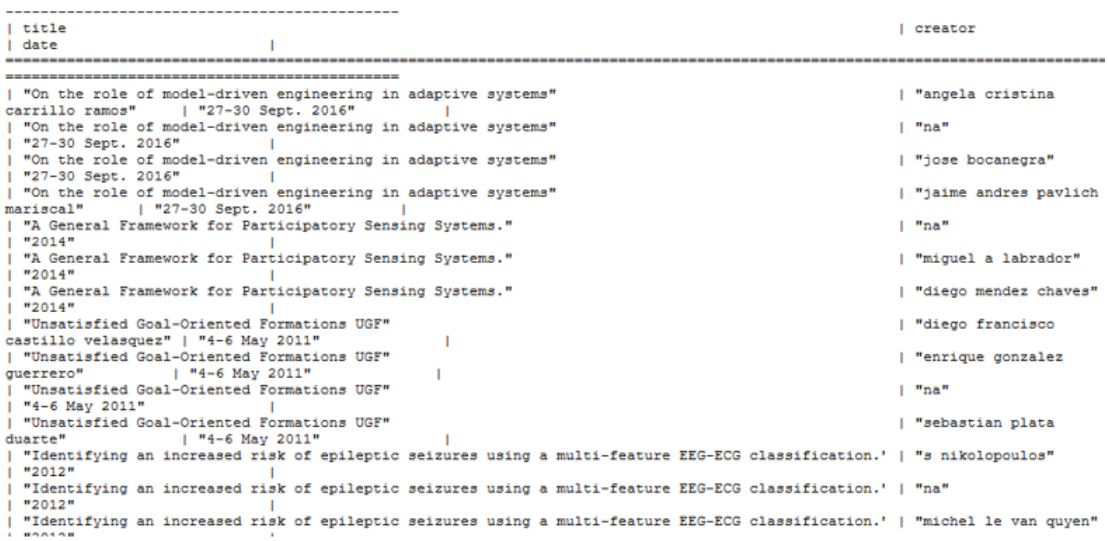


Figura 29. Consulta en el SPARQL *Endpoint* desplegado

En este caso se selecciona el servidor SPARQL *Fuseki* (ver Figura 29), dado que permite el acceso a través Internet y sirve de interfaz para la consulta por parte de aplicaciones que usen

[23] http://rdfpro.fbk.eu/

el protocolo HTTP. Una vez desplegado este servidor se procede a la carga del RDF fusionado, obtenido en la fase de generación. El despliegue de este servidor permite habilitar un punto de consulta (SPARQL *Endpoint*), donde se puede interactuar con la base de conocimiento conformada sobre autores y publicaciones científicas de la Facultad de Ingeniería de la Universidad Javeriana (ver Figura 30).



title	creator
"On the role of model-driven engineering in adaptive systems"	"angela cristina carrillo ramos"
"27-30 Sept. 2016"	
"On the role of model-driven engineering in adaptive systems"	"na"
"27-30 Sept. 2016"	
"On the role of model-driven engineering in adaptive systems"	"jose bocanegra"
"27-30 Sept. 2016"	
"On the role of model-driven engineering in adaptive systems"	"jaime andres pavlich mariscal"
"27-30 Sept. 2016"	
"A General Framework for Participatory Sensing Systems."	"na"
"2014"	
"A General Framework for Participatory Sensing Systems."	"miguel a labrador"
"2014"	
"A General Framework for Participatory Sensing Systems."	"diego mendez chaves"
"2014"	
"Unsatisfied Goal-Oriented Formations UGF"	"diego francisco castillo velasquez"
"4-6 May 2011"	
"Unsatisfied Goal-Oriented Formations UGF"	"enrique gonzalez guerrero"
"4-6 May 2011"	
"Unsatisfied Goal-Oriented Formations UGF"	"na"
"4-6 May 2011"	
"Unsatisfied Goal-Oriented Formations UGF"	"sebastian plata duarte"
"4-6 May 2011"	
"Identifying an increased risk of epileptic seizures using a multi-feature EEG-ECG classification.'	"s nikolopoulos"
"2012"	
"Identifying an increased risk of epileptic seizures using a multi-feature EEG-ECG classification.'	"na"
"2012"	
"Identifying an increased risk of epileptic seizures using a multi-feature EEG-ECG classification.'	"michel le van quyen"

Figura 30. Resultado retornado por el SPARQL *Endpoint*

Finalmente, la explotación se puede llevar acabo en cualquier aplicación o ambiente que pueda consumir el servicio REST expuesto por el *triple store* Fuseki a través del SPARQL *Endpoint*. A manera de ejemplo dentro de este proyecto se presentan algunas visualizaciones desarrolladas utilizando el lenguaje R, permitiendo analizar los resultados de la productividad científica de la comunidad de la Facultad de Ingeniería de la Universidad Javeriana de una forma más fácil e intuitiva. A continuación, se muestran algunas de las visualizaciones y métricas que se pueden obtener tras la realización de este trabajo.

1. Generación de nubes de palabras. A través de las palabras clave (*keywords*) presentes en los artículos publicados se construyen nubes de palabras para identificar los temas más frecuentemente abordados en dichas publicaciones (ver Figura 31). Estas nubes de palabras se pueden generar por autor, departamento o con toda la información disponible.

Figura 31. Nube de palabras

2. Gráfico de burbujas. En este tipo de visualización se pueden observar las métricas relacionadas con la producción de publicaciones, por año y por departamento (ver Figura 32). Asimismo, este gráfico se puede generar considerando las publicaciones de cada autor por año, permitiendo analizar la productividad científica a nivel de departamento o de cada uno de sus miembros.

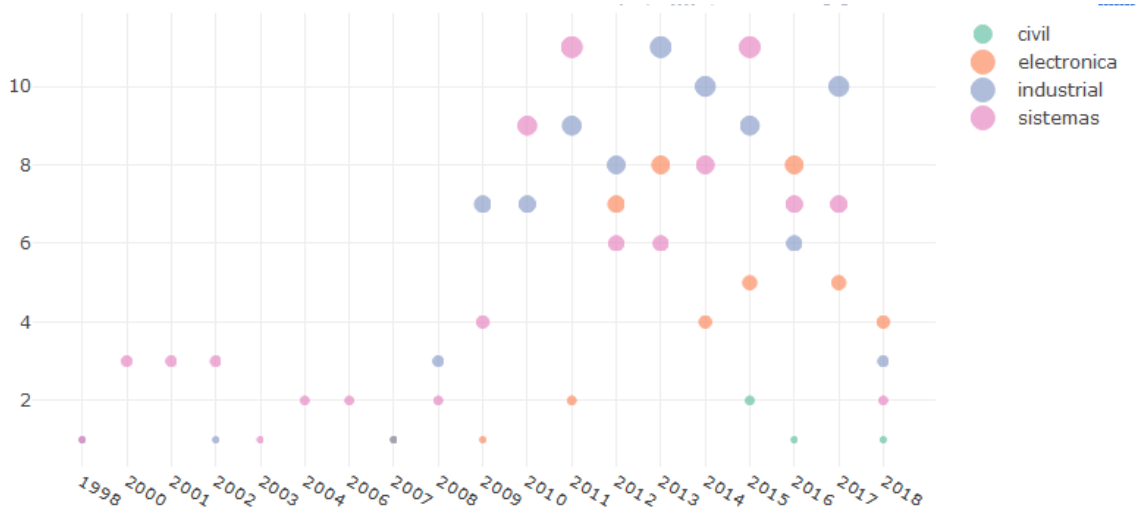


Figura 32. Gráfico de burbujas

3. Gráfico. La utilización de R permite generar diversos tipos de gráficos asociados a diferentes variables de los datos tratados. A modo de ejemplo, en la Figura 33 se muestra la participación de cada uno de los departamentos en relación con las publicaciones generadas por la Facultad, con lo cual, de una forma intuitiva se puede identificar el peso de cada departamento en cuanto a su producción científica.

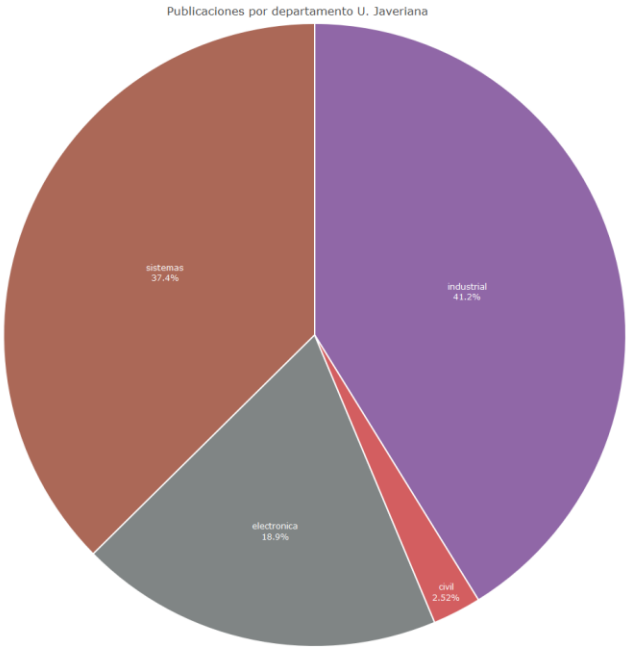


Figura 33. Gráfico de *pie*

4. Grafo de red. Este grafo permite explorar las relaciones existentes entre coautores y sus publicaciones. Por tanto, esta visualización permite identificar las redes de colaboración existentes entre los miembros de la comunidad javeriana por cada autor. A modo de ejemplo, en la Figura 34 se presenta el grafo de colaboración de la profesora Ángela Carrillo del departamento de sistemas de la Facultad de Ingeniería.

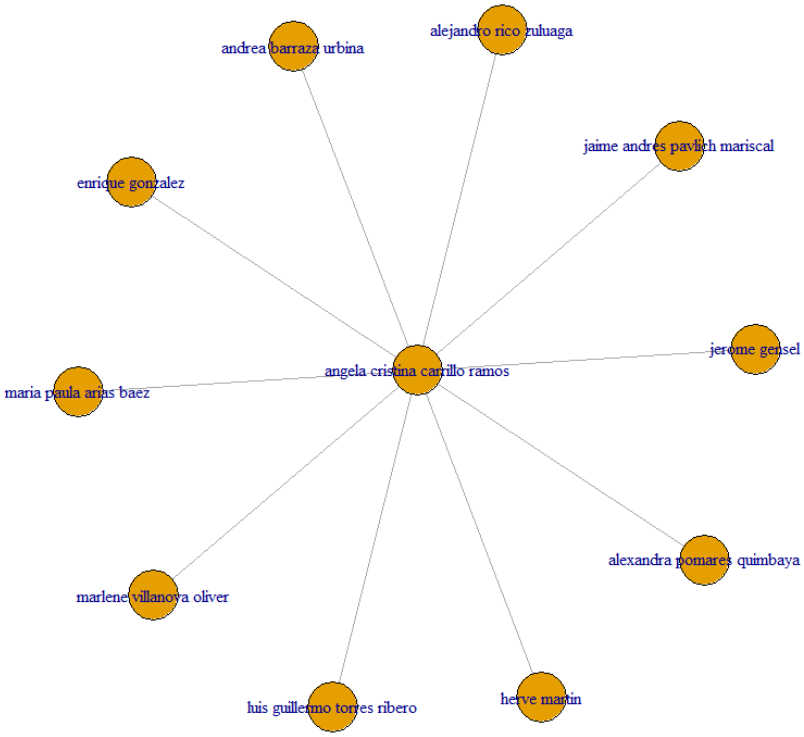


Figura 34. Red de colaboración de Ángela Carrillo

5. Gráfico de barras. Este resultado permite relacionar la cantidad de publicaciones con las ciudades en las que se publicaron. De esta manera, es posible identificar dónde se concentran los eventos científicos en los que participan y publican los miembros de la comunidad Javeriana (ver Figura 35).

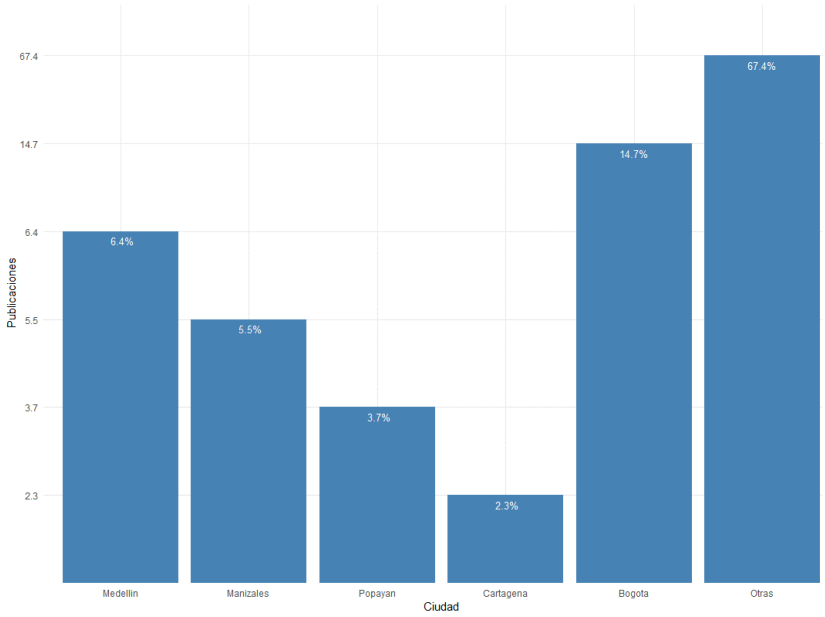


Figura 35. Gráfico de barras con las publicaciones por ciudad

VI. VALIDACIÓN

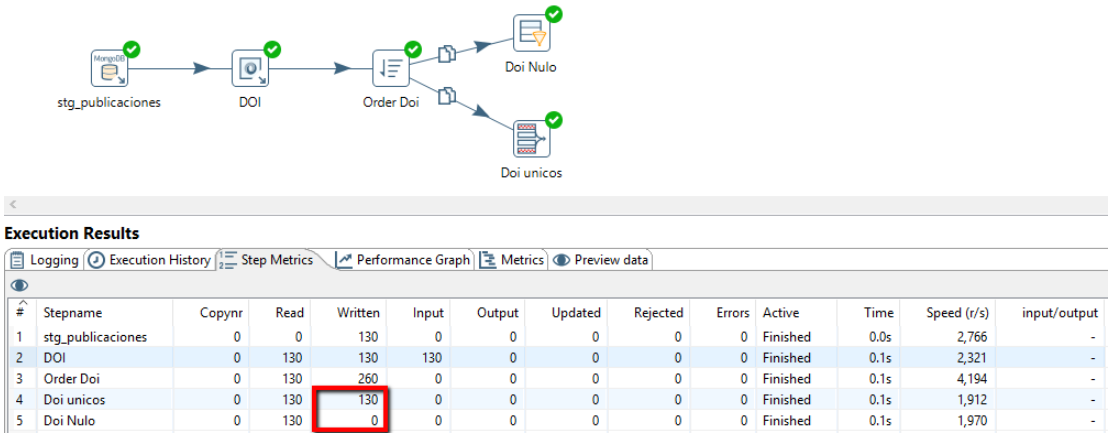
En esta sección se describen los procesos de validación realizados para los procesos de recuperación y transformación de datos a RDF, su publicación y explotación que representan el *core* del desarrollo de este trabajo.

Para las siguientes validaciones se saca una muestra de 130 publicaciones con 12 campos (ver Figura 36) para las que se ejecutan los siguientes procesos:

Figura 36. Datos sobre publicaciones

- Para la validación de la recuperación de datos se diseñan consultas contenidas en *workflows Pentaho* que permiten comprobar que la colección *Publicaciones* no presenta problemas relacionados con duplicados, registros nulos o publicaciones que no tienen un autor que haga parte del foco del proyecto, es decir, profesores de la Facultad de Ingeniería de la Universidad Javeriana.

  El análisis de la ejecución arroja resultados para los pasos de validación de 130 DOI únicos y 0 Registros con DOI nulo (ver Figura 37), siendo este el resultado esperado.



#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	stg_publicaciones	0	0	130	0	0	0	0	0	Finished	0.0s	2,766	-
2	DOI	0	130	130	130	0	0	0	0	Finished	0.1s	2,321	-
3	Order Doi	0	130	260	0	0	0	0	0	Finished	0.1s	4,194	-
4	Doi unicos	0	130	130	0	0	0	0	0	Finished	0.1s	1,912	-
5	Doi Nulo	0	130	0	0	0	0	0	0	Finished	0.1s	1,970	-

Figura 37. Validación de DOI nulos y DOI únicos

- La validación del RDF se adelanta por medio de consultas SPARQL cuyos resultados son comparados con el mapeo de las colecciones dentro de la base de datos no relacional y a partir de los cuales se puede verificar la completitud de los datos, conforme a los originales obtenidos de las diferentes API, una vez han sido transformados a formato RDF. En la Figura 38 se puede observar una consulta SPARQL almacenada en un objeto *Data Frame* (DF) de R, cuyo resultado son todas las propiedades que tiene una publicación (12) y limitado a los 10 primeros resultados, en caso de que alguna de las propiedades no pudiese ser consultada el servidor SPARQL retornaría error y no llegaría a llenarse el objeto DF.

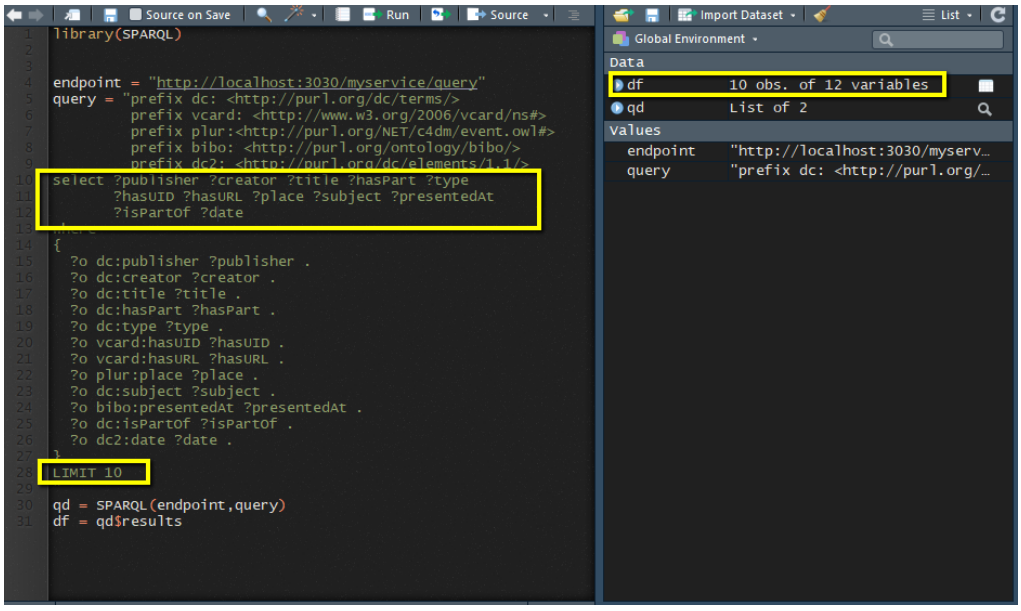


Figura 38. Validación, consulta publicación y sus propiedades.

- La publicación del RDF se valida a partir de conexiones HTTP al SPARQL *Endpoint* expuesto por el servidor *Fuseki*, comprobando que este responde correctamente a las consultas y permite acceder a los datos desde clientes como exploradores web o programas escritos en R.

- Finalmente, se valida la explotación gracias al desarrollo de programas en R que consumen los datos expuestos por el servidor *Fuseki* y que una vez leídos y almacenados en *Data Frames* pueden ser manipulados y analizados de forma similar a cualquier otro conjunto de datos (ver Figura 39).

publisher	creator	title	hasPart	type
dblp	juan arturo nolazco flores	Multi-speaker voice cryptographic key generation.	NA	Conference and Workshop Papers
dblp	juan arturo nolazco flores	Multi-speaker voice cryptographic key generation.	NA	Conference and Workshop Papers
dblp	l paola garcia perera	Multi-speaker voice cryptographic key generation.	NA	Conference and Workshop Papers
dblp	l paola garcia perera	Multi-speaker voice cryptographic key generation.	NA	Conference and Workshop Papers

Figura 39. Objeto *Data Frame* en R

VII. CONCLUSIONES, APORTES Y TRABAJO FUTURO

Conclusiones

El desarrollo de este proyecto es una aplicación de los principios de *Linked Open Data* en el que cada una de sus fases generó resultados que pueden ser utilizados en diferentes aplicaciones del contexto de publicaciones científicas. Gracias a la utilización de varias herramientas tecnológicas y lenguajes de programación fue posible generar valor en cada una de las fases y para el estudiante representó una experiencia valiosa que ha aportado en su formación, brindando la oportunidad de colaborar con participantes de proyectos similares y adquirir competencias en el proceso.

En el contexto de la metodología adoptada en este trabajo, durante la fase de especificación en la que se analizaron y seleccionaron las fuentes de información consideradas para alimentar el repositorio de publicaciones científicas generado, se encontraron diferentes problemas relacionados con la forma en la que se busca un autor dentro de las diferentes API tratadas. Estos problemas se relacionan con los caracteres especiales y las diversas formas en las que los nombres de los autores aparecen en las publicaciones, sobre todo, cuando se trabaja con nombres de origen latino. Este problema fue superado con el desarrollo de este trabajo.

En la fase de modelado se descubrieron varios recursos ontológicos en los que se encontraron ambigüedades en la descripción de los mismos, por lo que fue necesaria una constante revisión para asegurar la correcta descripción de los datos. Teniendo en cuenta estas ambigüedades, resulta importante revisar las fuentes directas de las ontologías y no solo guiarse por los resultados de repositorios de recursos ontológicos. También resulta fundamental considerar el contexto de los recursos ontológicos, con el fin de comprender la semántica de los diversos elementos que conforman los distintos recursos y que, incluso, pueden tener el mismo nombre.

La generación del RDF, cuyo insumo es la red de ontologías desarrollada y el mapeo de los datos, requirió del uso de diferentes herramientas que presentaron diversos desafíos. Los *plugins* de *Pentaho* que hacen parte de la plataforma LOD-GF requieren de configuraciones que deben hacerse a bajo nivel, llegando a requerir modificaciones de las configuraciones de los mismos, debido a que los desarrollos asociados a esta herramienta requieren ajustes para ambientes Windows.

Con la implementación del servidor SPARQL *Fuseki* fue posible desplegar la base de conocimiento producto de la integración semántica construida con la información de los diferentes repositorios considerados. Esto permite desplegar una interfaz de consulta a los datos en RDF a través de un SPARQL *Endpoint*, donde los datos se encuentras integrados y disponibles para su consulta.

En relación a la explotación de los datos, se lograron generar diversos ejemplos desarrollados en R para demostrar el potencial de aplicación que tiene este proyecto para análisis bibliométricos, gracias al uso del insumo representado por el repositorio de conocimiento científico conforme a los principios de *Linked Data* entregado en este proyecto.

Aportes

Algunos de los aportes más significativos de este trabajo son:

- Diseño de una arquitectura que soporta la aplicación de los principios de *Linked Open Data*.
- Desarrollo de una propuesta de integración semántica y despliegue de una base de conocimiento con las publicaciones científicas de los departamentos de la Facultad de Ingeniería de la Universidad Javeriana.
- Desarrollo de una herramienta de generación de datos para análisis y estadística que puede ser explotada por cualquier Facultad o departamento de la Universidad Javeriana.
- Disposición de un SPARQL *Endpoint* a la comunidad académica con los datos de publicaciones científicas generadas desde la Facultad de Ingeniería de la Universidad Javeriana.
- Desarrollo de procesos de extracción y control de consultas masivas en la herramienta *Pentaho* que pueden ser reutilizados en todo tipo de proyectos.

Trabajos futuros

Se propone como trabajo futuro la incorporación de nuevas fuentes de información que pueden ser repositorios que requieren suscripción y pago para extender las capacidades del repositorio y posibles nuevas funcionalidades para el repositorio construido, como por ejemplo el análisis de cocitación o de relevancia/impacto de las publicaciones.

Además, se trabajará en el agendamiento de todos los *workflows* desarrollados en *Pentaho*, de forma que se pueda formalizar un proceso de carga inicial y mantenimiento del repositorio que asegure su relevancia en el tiempo.

Por otro lado, la explotación de los datos representa el paso natural a seguir en el desarrollo de este trabajo, brindándole acceso a los datos a personas expertas en analítica se podrían identificar aplicaciones que no han sido contempladas por parte de este proyecto, entre ellas, el desarrollo en profundidad de un análisis bibliométrico.

Finalmente, se trabajará en el desarrollo de un *dashboard* que facilite la utilización de los datos entregados como soporte para la toma de decisiones administrativas, ya que a través de estos datos se pueden llevar a cabo políticas de fomento a la investigación en ciertos temas, o bien facilitar la conexión de ciertos autores que trabajan en temas similares, que no han colaborado en la elaboración de publicaciones científicas.

REFERENCIAS

[1] «Sci2 Tool,» [En línea]. Available: https://sci2.cns.iu.edu/user/index.php.

[2] «CiteSpace,» [En línea]. Available: http://cluster.cis.drexel.edu/~cchen/citespace/.

[3] C. Bizer, T. Heath y T. Berners-Lee, «Linked Data: The Story So Far,» *International Journal on Semantic Web and Information Systems,* 2011.

[4] Y. Hu, K. Janowicz, G. McKenzie, K. Sengupt y P. Hitzler, «A Linked-Data-driven and Semantically-enabled Journal Portal for Scientometrics,» *The Semantic Web – ISWC 2013,* 2013.

[5] C. Bizer, T. Heath y T. Berners-Lee, «Linked Data: The Story So Far,» *International Journal on Semantic Web and Information Systems,* 2011.

[6] M. Hausenblas, Linked Data Applications, Galway, Ireland: DIGITAL ENTERPRISE RESEARCH INSTITUTE, 2009.

[7] «W3C,» [En línea]. Available: https://www.w3.org/standards/semanticweb/data.

[8] A. Latif, M. Tanvir Afzal, D. Helic, K. Tochtermann y H. Maurer, «Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal),» 2010.

[9] «W3,» 208. [En línea]. Available: https://www.w3.org/TR/rdf-sparql-query/.

[10] L. M. V. O. C. &. A. G.-P. Boris Villazón-Terrazas, «Methodological Guidelines for Publishing Government Linked Data,» *Linking Government Data,* pp. 27-49, 2011.

[11] «W3,» 2016. [En línea]. Available: https://www.w3.org/wiki/LinkedData.

[12] A. Gómez Pérez y M. C. Suárez-Figueroa, «NeOn Methodology for Building Ontology Networks: a,» 2009.

[13] T. Berners-Lee, Y. Che, L. Chilton, D. Connolly, . R. Dhanaraj, J. Hollenbach, A. Lerer y D. Sheets, «Tabulator: Exploring and Analyzing linked data,» 2006.

[14] A. Gerber y A. V. d. Merwe, «A Functional Semantic Web Architecture,» 2008.

[15] D. Fensel y F. Facca, «Semantic Web Architecture,» 2010.

[16] R. W. Group, «W3C Semantic Web,» [En línea]. Available: https://www.w3.org/RDF/.

[17] M. Egaña, «Datos Enlazados y Web Semántica».

[18] W3C, «W3C Recomendation,» 2014. [En línea]. Available: https://www.w3.org/TR/rdf-schema/#ch_introduction.

[19] W3C, «SPARQL Query Language for RDF,» 2008. [En línea]. Available: https://www.w3.org/TR/rdf-sparql-query/#introduction.

[20] C. Chua, L. Cao, K. Cousins, K. Mohan y D. Straub, «IS Bibliuographic Repository (ISBIB): A Central Repository of Reseach Information for the IS Community,» *Communications of the Association for Informacion Systems*, 2002.

[21] «Colorado Alliance of Reseach Libraries,» [En línea]. Available: https://www.coalliance.org/about.

[22] «Clarivate Analytics,» [En línea]. Available: https://clarivate.com/products/endnote/. [Último acceso: 2018].

[23] «Colciencias,» [En línea]. Available: https://scienti.colciencias.gov.co/cvlac/Login/pre_s_login.do. [Último acceso: 2018].

[24] «SCi2 Tool,» [En línea]. Available: https://sci2.cns.iu.edu/user/index.php. [Último acceso: 2018].

[25] «CiteSpace,» [En línea]. Available: http://cluster.cis.drexel.edu/~cchen/citespace/. [Último acceso: 2018].

[26] O. Seneviratne, v. W. Patton, D. Miao, F. Shih, W. Li, L. Kagal y C. Castillo, «Developing Mobile Linked Data Applications».

[27] P. Frischmuth, J. Klímek, S. Auer, S. Tramp, J. Unbehauen, K. Holzweißig y C.-M. Marquardt, «Linked Data in Enterprise Information,» 2012.

[28] T. Primov, «OntoText,» 2016. [En línea]. Available: https://www.ontotext.com/linked-data-solutions-in-healthcare/.

[29] «Scrapy,» [En línea]. Available: https://scrapy.org/.

[30] C. Kreibich, «GitHub,» 2014. [En línea]. Available: https://github.com/ckreibich/scholar.py.

[31] «IEEE: Institute of Electrical and Electronics Engineers,» 2018. [En línea]. Available: https://developer.ieee.org/docs.

[32] DBLP computer science bibliography, 2018. [En línea]. Available: https://dblp.uni-trier.de/faq/13501473.

[33] «dblp computer science bibliography,» [En línea]. Available: http://dblp.org/faq/. [Último acceso: 2018].

[34] Semantic Scholar, [En línea]. Available: https://api.semanticscholar.org/.

[35] «Semantic Scholar,» [En línea]. Available: https://www.semanticscholar.org/faq. [Último acceso: 2018].

[36] R. Iannella y J. McKinney, «W3C Interest Group,» [En línea]. Available: https://www.w3.org/TR/vcard-rdf/.

[37] B. D'Arcus y F. Giasson, «The Bibliographic Ontology,» 2008. [En línea]. Available: http://bibliontology.com/.

[38] D. U. Board, «Dublin Core Metadata initiative,» 2012. [En línea]. Available: http://dublincore.org/documents/dcmi-terms/.

[39] D. Brickley y L. Miller, «FOAF Vocabulary Specification,» 2014. [En línea]. Available: http://xmlns.com/foaf/spec/.

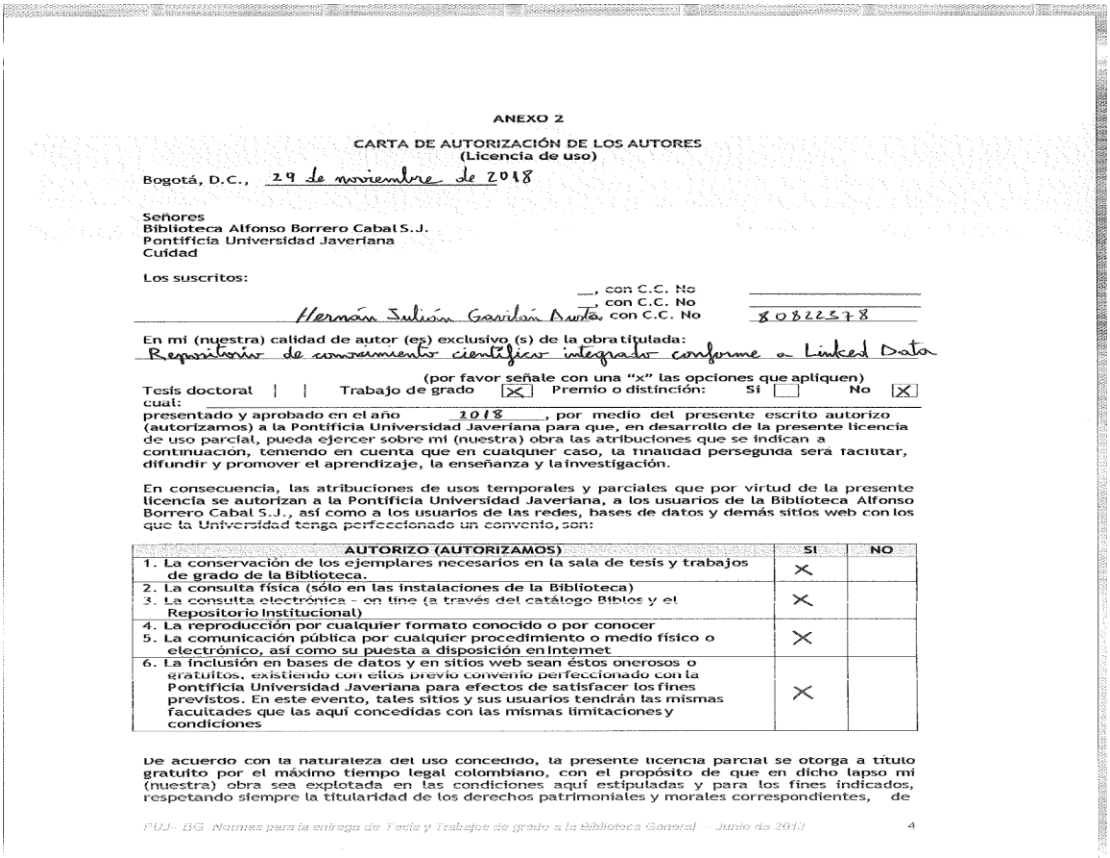
[40] I. HUSSAIN y S. ASGHAR, «A survey of author name disambiguation techniques:,» 2017.

[41] O. W. Group, «W3C Semantic Web,» 2012. [En línea]. Available: https://www.w3.org/OWL/.

[42] C. R. Hernández Sampieri, C. Fernández Collado y P. Baptista Lucio, Metodología de la Investigación, 2006.

[43] O. Stephens, «Selecting with SPARQL,» 2013.

ANEXOS



ANEXO 2

CARTA DE AUTORIZACIÓN DE LOS AUTORES
(Licencia de uso)

Bogotá, D.C., 29 de noviembre de 2018

Señores
Biblioteca Alfonso Borrero Cabal S.J.
Pontificia Universidad Javeriana
Cuidad

Los suscritos:

__, con C.C. No
__, con C.C. No
Hermán Julián Garzón Ruiz, con C.C. No 80822578

En mi (nuestra) calidad de autor (es) exclusivo (s) de la obra titulada:
Repositorio de conocimiento científico integrado conforme a Linked Data

(por favor señale con una "x" las opciones que apliquen)
Tesis doctoral | | Trabajo de grado [X] Premio o distinción: Si [ ] No [X]
cual: _____
presentado y aprobado en el año 2018, por medio del presente escrito autorizo (autorizamos) a la Pontificia Universidad Javeriana para que, en desarrollo de la presente licencia de uso parcial, pueda ejercer sobre mi (nuestra) obra las atribuciones que se indican a continuación, teniendo en cuenta que en cualquier caso, la finalidad perseguida sera facilitar, difundir y promover el aprendizaje, la enseñanza y la investigación.

En consecuencia, las atribuciones de usos temporales y parciales que por virtud de la presente licencia se autorizan a la Pontificia Universidad Javeriana, a los usuarios de la Biblioteca Alfonso Borrero Cabal S.J., así como a los usuarios de las redes, bases de datos y demás sitios web con los que la Universidad tenga perfeccionado un convenio, son:

AUTORIZO (AUTORIZAMOS)	SI	NO
1. La conservación de los ejemplares necesarios en la sala de tesis y trabajos de grado de la Biblioteca.	X	
2. La consulta física (sólo en las instalaciones de la Biblioteca)		
3. La consulta electrónica - on line (a través del catálogo Biblos y el Repositorio Institucional)	X	
4. La reproducción por cualquier formato conocido o por conocer		
5. La comunicación pública por cualquier procedimiento o medio físico o electrónico, así como su puesta a disposición en Internet	X	
6. La inclusión en bases de datos y en sitios web sean éstos onerosos o gratuitos, existiendo con ellos previo convenio perfeccionado con la Pontificia Universidad Javeriana para efectos de satisfacer los fines previstos. En este evento, tales sitios y sus usuarios tendrán las mismas facultades que las aquí concedidas con las mismas limitaciones y condiciones	X	

De acuerdo con la naturaleza del uso concedido, la presente licencia parcial se otorga a título gratuito por el máximo tiempo legal colombiano, con el propósito de que en dicho lapso mi (nuestra) obra sea explotada en las condiciones aquí estipuladas y para los fines indicados, respetando siempre la titularidad de los derechos patrimoniales y morales correspondientes, de

PUJ- BG Normas para la entrega de Tesis y Trabajos de grado a la Biblioteca General - Junio de 2013 4

acuerdo con los usos honrados, de manera proporcional y justificada a la finalidad perseguida, sin ánimo de lucro ni de comercialización.

De manera complementaria, garantizo (garantizamos) en mi (nuestra) calidad de estudiante (s) y por ende autor (es) exclusivo (s), que la Tesis o Trabajo de Grado en cuestión, es producto de mi (nuestra) plena autoría, de mi (nuestro) esfuerzo personal intelectual, como consecuencia de mi (nuestra) creación original particular y, por tanto, soy (somos) el (los) único (s) titular (es) de la misma. Además, aseguro (aseguramos) que no contiene citas, ni transcripciones de otras obras protegidas, por fuera de los límites autorizados por la ley, según los usos honrados, y en proporción a los fines previstos; ni tampoco contempla declaraciones difamatorias contra terceros; respetando el derecho a la imagen, intimidad, buen nombre y demás derechos constitucionales. Adicionalmente, manifiesto (manifestamos) que no se incluyeron expresiones contrarias al orden público ni a las buenas costumbres. En consecuencia, la responsabilidad directa en la elaboración, presentación, investigación y, en general, contenidos de la Tesis o Trabajo de Grado es de mí (nuestro) competencia exclusiva, eximiendo de toda responsabilidad a la Pontificia Universidad Javeriana por tales aspectos.

Sin perjuicio de los usos y atribuciones otorgadas en virtud de este documento, continuaré (continuaremos) conservando los correspondientes derechos patrimoniales sin modificación o restricción alguna, puesto que de acuerdo con la legislación colombiana aplicable, el presente es un acuerdo jurídico que en ningún caso conlleva la enajenación de los derechos patrimoniales derivados del régimen del Derecho de Autor.

De conformidad con lo establecido en el artículo 30 de la Ley 23 de 1982 y el artículo 11 de la Decisión Andina 351 de 1993, *"Los derechos morales sobre el trabajo son propiedad de los autores"*, los cuales son irrenunciables, imprescriptibles, inembargables e inalienables. En consecuencia, la Pontificia Universidad Javeriana está en la obligación de RESPETARLOS Y HACERLOS RESPETAR, para lo cual tomará las medidas correspondientes para garantizar su observancia.

NOTA: Información Confidencial:
Esta Tesis o Trabajo de Grado contiene información privilegiada, estratégica, secreta, confidencial y demás similar, o hace parte de una investigación que se adelanta y cuyos resultados finales no se han publicado. Si [ ] No [X]
En caso afirmativo expresamente indicaré (indicaremos), en carta adjunta, tal situación con el fin de que se mantenga la restricción de acceso.

NOMBRE COMPLETO	No. del documento de identidad	FIRMA
Hernán Julián Gavilán Acosta	80822378	

FACULTAD: Ingenierías
PROGRAMA ACADÉMICO: Maestría en Ingeniería de Sistemas y Computación

PUJ– BG Normas para la entrega de Tesis y Trabajos de grado a la Biblioteca General — Junio de 2013 5

ANEXO 3
BIBLIOTECA ALFONSO BORRERO CABAL, S.J.
DESCRIPCIÓN DE LA TESIS O DEL TRABAJO DE GRADO
FORMULARIO

TÍTULO COMPLETO DE LA TESIS DOCTORAL O TRABAJO DE GRADO			
Repositorio de conocimiento científico integrado conforme a Linked Data			
SUBTÍTULO, SI LO TIENE			
AUTOR O AUTORES			
Apellidos Completos		**Nombres Completos**	
Gaitán Ariza		Hernán Julián	
DIRECTOR (ES) TESIS O DEL TRABAJO DE GRADO			
Apellidos Completos		**Nombres Completos**	
Vilches Blázquez		Luis Manuel	
FACULTAD			
Ingeniería			
PROGRAMA ACADÉMICO			
Tipo de programa (seleccione con "x")			
Pregrado	**Especialización**	**Maestría**	**Doctorado**
		X	
Nombre del programa académico			
Maestría en Ingeniería de Sistemas y Computación			
Nombres y apellidos del director del programa académico			
Angela Cristina Carrillo Ramos			
TRABAJO PARA OPTAR AL TÍTULO DE:			
Magíster en Ingeniería de Sistemas y Computación			
PREMIO O DISTINCIÓN *(En caso de ser LAUREADAS o tener una mención especial)*:			

CIUDAD	AÑO DE PRESENTACIÓN DE LA TESIS O DEL TRABAJO DE GRADO	NÚMERO DE PÁGINAS
Bogotá	2018	79

TIPO DE ILUSTRACIONES (seleccione con "x")						
Dibujos	**Pinturas**	**Tablas, gráficos y diagramas**	**Planos**	**Mapas**	**Fotografías**	**Partituras**
		X				

SOFTWARE REQUERIDO O ESPECIALIZADO PARA LA LECTURA DEL DOCUMENTO
Nota: En caso de que el software (programa especializado requerido) no se encuentre licenciado por la Universidad a través de la Biblioteca (previa consulta al estudiante), el texto de la Tesis o Trabajo de Grado quedará solamente en formato PDF.

PUJ - BG Normas para la entrega de Tesis y Trabajos de grado a la Biblioteca General — Junio de 2013 6

MATERIAL ACOMPAÑANTE					
TIPO	DURACIÓN (minutos)	CANTIDAD	FORMATO		
			CD	DVD	Otro ¿Cuál?
Vídeo					
Audio					
Multimedia					
Producción electrónica					
Otro Cuál?					

DESCRIPTORES O PALABRAS CLAVE EN ESPAÑOL E INGLÉS

Son los términos que definen los temas que identifican el contenido. (*En caso de duda para designar estos descriptores, se recomienda consultar con la Sección de Desarrollo de Colecciones de la Biblioteca* Alfonso Borrero Cabal S.J *en el correo* biblioteca@javeriana.edu.co, *donde se les orientará).*

ESPAÑOL	INGLÉS
Linked Data-	Linked Data-
Web Semantico-	Semantic Web
Repositorio-	Repository
Publicacion Cientifica-	Scientific publication

RESUMEN DEL CONTENIDO EN ESPAÑOL E INGLÉS
(Máximo 250 palabras - 1530 caracteres)

Memoria del trabajo de grado y anexos

PUJ- BG Normas para la entrega de Tesis y Trabajos de grado a la Biblioteca General – Junio de 2013 7