



Palabra Clave (La Plata)  
ISSN: 1853-9912  
palabraclave@fahce.unlp.edu.ar  
Universidad Nacional de La Plata  
Argentina

# Integración semiautomática de tecnologías de la web semántica en bases de datos de patentes

**Reyes-Lillo, Danilo**

Integración semiautomática de tecnologías de la web semántica en bases de datos de patentes

Palabra Clave (La Plata), vol. 8, núm. 1, 2018

Universidad Nacional de La Plata, Argentina

**Disponible en:** <https://www.redalyc.org/articulo.oa?id=350556354003>

**DOI:** <https://doi.org/10.24215/18539912e057>



Esta obra está bajo una Licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional.

## Integración semiautomática de tecnologías de la web semántica en bases de datos de patentes

Semi-automatic integration of semantic web technologies in patent databases

Daniilo Reyes-Lillo

Universidad de Playa Ancha. Facultad de Ciencias Sociales.

Departamento de Ciencias de la Documentación, Chile

danilo.reyes@upla.cl

DOI: <https://doi.org/10.24215/18539912e057>

Redalyc: <https://www.redalyc.org/articulo.oa?id=350556354003>

Recepción: 24 Abril 2018

Aprobación: 22 Mayo 2018

### RESUMEN:

Se analiza la aplicación de tecnologías de la web semántica al ámbito de las patentes y se propone una estrategia metodológica de integración semiautomática de tecnologías de *linked data* a bases de datos relacionales de patentes. En primera instancia, se conceptualiza la documentación de patentes y su potencial integración de tecnologías de web semántica. A continuación, se analiza el modelo de datos para establecer las relaciones potenciales de la fuente de datos con una ontología del ámbito de la propiedad industrial. Posteriormente, se efectúa un *mapping* o una traducción de la base de datos a RDF basado en R2RML mediante Karma, un *software* de integración de datos. Finalmente, se describe el procedimiento de modelado de datos para la publicación de información de patentes en un modelo RDF.

**PALABRAS CLAVE:** Web semántica, Linked Data, Bases de datos de patentes, Búsqueda semántica de patentes, Mapping.

### ABSTRACT:

The application of semantic web technologies to the field of patents is analyzed and a semi-automatic integration methodology of linked data technologies to patent relational databases is proposed. At first, patent documentation and its potential integration of semantic web technologies are conceptualized. Next, the data model is analyzed to establish the potential relationships of the data source with an ontology of the field of industrial property. Subsequently, a mapping or a translation of the database to RDF based on R2RML is carried out using Karma, a data integration software. Finally, the data modeling procedure for the publication of patent information in an RDF model is described.

**KEYWORDS:** Semantic Web, Linked Data, Patent Databases, Semantic Patent Search, Mapping.

### INTRODUCCIÓN

La integración de tecnologías de la web semántica en distintos servicios de información es una realidad que ha ido creciendo y se ha transformado en un desafío para muchos organismos e instituciones a nivel mundial que ofrecen sus datos de forma abierta. Asimismo, el interés por el mundo de la propiedad industrial y las patentes de invención también resulta creciente debido a su utilidad como fuente de información fundamental para promover la investigación, desarrollo e innovación (I+D+i). Sin embargo, aunque la Organización Mundial de Propiedad Intelectual (OMPI) se refiere a la existencia de servicios de búsqueda semántica en bases de datos de patentes (OMPI, 2016), estos no están relacionados con las tecnologías de la web semántica o *Linked Data* (LD). Por otro lado, si bien desde la óptica de la investigación científica de orden más teórico se ha empezado a incursionar en un potencial vínculo entre la web semántica y la búsqueda de patentes, en general, no han terminado por converger en servicios de búsqueda semántica concretos.

El mencionado vínculo potencial se plasma en la posibilidad que otorgan las tecnologías de la web semántica para optimizar la estructura de la información en base a ontologías, enriquecer los datos y facilitar la reutilización por parte de actores interesados en la propiedad industrial, como empresas, centros de investigación y otros actores (García Moreno, 2015).

Bajo esa perspectiva, se propone determinar una estrategia procedimental para la integración de tecnologías semánticas a una base de datos de patentes cimentada en una estructura de base de datos relacional, a través de un proceso denominado *mapping*. La finalidad es fortalecer la integración de datos y a su vez proporcionar al servicio la posibilidad de enlazar y de que sus datos sean enlazados por otras fuentes que hagan más potentes sus registros, a través de un portal de datos enlazados. Lo anterior promueve un mejor acceso y comprensión de la información de las patentes para las empresas, centros de investigación, universidades, inventores y otros actores que intervienen en la I+D+i.

## PATENTES Y SU RELEVANCIA COMO FUENTE DE DATOS

La investigación, el desarrollo y la innovación (I+D+i) son un pilar fundamental para el Estado, las universidades, los centros de investigación y las empresas (Ertzkowitz y Leydesdorff, 2000). En este contexto, el acceso y la utilización de información científico-tecnológica se transforma en un elemento sustancial para las personas y organizaciones que buscan generar nuevo conocimiento o nuevos desarrollos tecnológicos. En ese plano, la propiedad intelectual toma relevancia, ya que es un cuerpo legal que se refiere a la protección de todas las creaciones del intelecto humano. En el convenio que establece la creación de la OMPI se indica la materia que puede ser susceptible de protección mediante derechos de propiedad intelectual:

- Obras literarias, artísticas y científicas.
- Interpretaciones y ejecuciones de obras, junto con fonogramas.
- Invenciones de cualquier campo de la actividad humana.
- Descubrimientos científicos.
- Diseños industriales.
- Marcas y denominaciones comerciales y no comerciales.
- Protección contra competencia desleal.
- Todos los derechos relativos a la actividad intelectual del hombre, en materia industrial, artística, literaria y/o científica.

Si profundizamos en el área de la documentación tecnológica, la propiedad industrial, como rama de la propiedad intelectual, protege invenciones de aplicación industrial en base a patentes, modelos de utilidad, diseños industriales, topografías de circuitos integrados, indicaciones geográficas y denominaciones de origen (OMPI, 2016). En este ámbito, las patentes de invención toman especial relevancia, debido a que el sistema de patentes tiene como finalidad incentivar la innovación y el crecimiento económico mediante dos mecanismos:

1. Entrega protección a la creatividad y una recompensa para los inventores a raíz de las fuertes inversiones efectuadas en una nueva invención. Las patentes protegen por un período de tiempo determinado y de manera territorial los derechos del inventor sobre su desarrollo.
2. Publicación y divulgación de toda la información técnica referida a una nueva invención a través de un documento de patente. Dicho documento se publica luego de un plazo de tiempo determinado, una vez presentada la solicitud y da cuenta de todos los detalles técnicos de la invención para que la comunidad lo pueda consultar (OMPI, 2013).

Asimismo, la protección legal que otorga una patente es de carácter territorial, es decir, se limita a un determinado país o grupo de países. Sin embargo, la información técnica que divulga la patente se visibiliza de manera mundial, es decir, se puede consultar desde cualquier lugar del mundo. Si la patente no está protegida en un país, se considera de dominio público.

Las patentes son fuentes de gran relevancia en materia de información científico-técnica, legal y comercial, por ende, resultan un recurso fundamental para investigadores, emprendedores e inventores. Adicionalmente, son documentos que poseen información exclusiva y original, por lo tanto, no ha sido publicada previamente en artículos científicos, *papers* de conferencias u otro tipo de documentación de naturaleza similar. Se estima que alrededor del 70 u 80 por ciento de la información contenida en una patente no ha sido publicado en ningún otro lugar (Singh, Chakraborty y Vincent, 2016). Por otro lado, a causa de la norma ST.9 de la OMPI, que define los metadatos bibliográficos fundamentales que deben estar contenidos en un documento de patente, es que la información que ofrecen dichos documentos está altamente estructurada y normalizada, lo que facilita la posibilidad de hacer análisis y procesamiento masivo de patentes.

Todo lo anterior hace de las patentes un recurso de información altamente atractivo, no sólo para científicos o desarrolladores de invenciones tecnológicas, sino también para otros actores involucrados en el entorno de la propiedad industrial como abogados especialistas en propiedad intelectual, especialistas de *marketing* o estrategias comerciales, y especialistas en transferencia tecnológica.

Con base en lo anterior, es posible afirmar que el objetivo de las búsquedas para el uso de patentes va a depender de la naturaleza de la necesidad y el perfil que tenga el usuario final. En ese sentido, es posible tipificar las búsquedas considerando los factores previamente señalados (Hunt, Nguyen y Rodgers, 2012). De esta manera, es factible identificar los siguientes tipos de búsqueda de patentes:

- Búsqueda de patentabilidad: se utiliza para determinar la posibilidad de patentar una invención, considerando elementos como la novedad (que no exista con anterioridad en el estado de la técnica) o su nivel inventivo (que no se derive de manera evidente a partir del conocimiento preexistente).
- Búsqueda de infracción: se utiliza para determinar si una patente vigente afirma lo mismo que determinado concepto o invento no patentado. Para esta búsqueda se incluyen sólo patentes que no han expirado.
- Búsqueda de autorización (búsqueda de derecho a uso o libertad de operar): se utiliza para determinar si una invención tiene autorización legal para ser utilizada, comercializada o replicada. Dicha autorización se establece cuando una patente no ha sido infringida o ha expirado.
- Búsqueda de estado del arte o estado de la técnica: es una búsqueda exhaustiva de todas las patentes disponibles y documentos no patentados, sin enfocarse sólo en un invento, sino que se reúne todo lo relacionado a un campo técnico. Todos estos documentos reunidos reflejan el “estado de arte” o “estado de la técnica”. La extensión y profundidad de la búsqueda dependerá de qué tan acotada fue la definición de la tecnología.
- Búsqueda de actividad de patentamiento (*patent landscape*): es un análisis más profundo de las invenciones, patentadas o no, después de la realización del estado de arte. Este estudio resulta –en la categorización de patentes– en descubrimientos fundamentales versus mejoras incrementales, una visualización de las patentes en el tiempo, la historia del desarrollo de la tecnología, e incluso el análisis de las colaboraciones o citas de inventores.

Con base en la categorización de búsqueda de patentes antes expuesta, es posible concluir que cada búsqueda de documentación científico-tecnológica dependerá del objetivo del usuario. Sin embargo, las patentes como fuente de información resultan útiles para diferentes objetivos, ya sea desde el punto de vista legal, económico, comercial o inventivo. Una patente es un documento que ofrece muchos datos utilizables en distintos contextos.

A pesar de lo anterior, la naturaleza técnica y jurídica de las patentes hace muy compleja no sólo su lectura, sino también las maneras de buscar y acceder a ellas (Adams, 2012). Por este motivo, se plantea la integración de tecnologías de la web semántica o *linked data* como un mecanismo que facilite la obtención de información tecnológica en bases de datos de patentes.

## LA WEB SEMÁNTICA Y MECANISMOS DE INTEGRACIÓN DE DATOS ENLAZADOS

La web semántica no es un concepto nuevo, sino que fue la idea original de Tim Berners-Lee, el inventor de la web, quien planteó la necesidad de agregar el concepto semántica a la web (Berners-Lee et al., 1994). La aplicación de esta idea queda definida como una extensión de la actual web que se caracteriza por entregar un significado bien definido para que las máquinas y las personas puedan comprenderlo. Se basa en la idea de que, mediante datos definidos y enlazados aplicaciones heterogéneas puedan encontrar, integrar y reutilizar datos de la web.

Con base en lo anterior, la web semántica se caracteriza como una extensión de la web actual, es decir, coexiste con todas las tecnologías de la web. Asimismo, resalta la necesidad de definir los datos para entregarle significado (semántica). Finalmente, declara la opción de que las máquinas sean capaces de manipular y trabajar con los datos (Pastor-Sánchez, 2011).

De esta manera, a los elementos, lenguajes y protocolos tradicionales de la web (HTML, CSS, URLs, entre otros) es necesario adicionarles una serie de elementos tecnológicos involucrados que dan vida a la web semántica como los metadatos, vocabularios, ontologías, elementos identificadores y relaciones entre los datos. Dichos elementos constituyen una arquitectura de la web semántica (Gerber, Barnard y Van, 2007).

Dentro de los múltiples elementos que componen la arquitectura de la web semántica, se destaca el rol de las URIs, las ontologías y el lenguaje RDF.

Las URIs o Identificadores de Recursos Uniformes (*Uniform Resource Identifier*) fortalecen la identificación y localización de recursos de información en el contexto de la web semántica. Una URI se define como una cadena de caracteres que posibilita, por un lado, la identificación de recursos de información localizables en la web y, por otro, identifica entidades concretas o abstractas que no se transmiten vía web, por ejemplo, un individuo, un artefacto o un sentimiento. Estos identificadores resultan fundamentales para la web semántica porque se busca transformar la web sintáctica a una web en la que la unidad mínima de información sea un dato. En este marco, cada dato se identifica con una URI y es posible establecer relaciones con otros datos, que a su vez estarán identificados con otras URIs. Incluso las relaciones pueden estar identificadas con una URI (Labra Gayo, 2011).

*Resource Description Framework* (RDF) es un lenguaje que permite identificar y describir recursos, sus propiedades y relaciones con otros recursos mediante tripletas (sujeto - predicado - objeto). Dicho modelo resulta también adecuado para representar datos (Decker et al., 2000), ya que ofrece una manera minimalista de describir los recursos de información en la web bajo la estructura de las tripletas (Shadbolt, Hall y Berners-Lee, 2006).

Por su parte, las ontologías buscan describir los objetos y sus relaciones definiendo clases, subclases, propiedades y subpropiedades. Normalmente las ontologías funcionan con OWL (*Web Ontology Language*) que complementa a RDF expresando las relaciones semánticas. En este marco, las ontologías resultan fundamentales en el ámbito de la web semántica, ya que cumplen el rol de mantener las semánticas bien definidas estructurando y ejecutando los procesos de inferencia (Pastor-Sánchez, 2011). Aunque RDF presenta elementos semánticos básicos, los que ofrecen las ontologías son más complejos y avanzados porque posee mayor capacidad para interpretar el contenido por parte de máquinas que en relación a otros formatos.

## INTEGRACIÓN DE TECNOLOGÍAS DE LA WEB SEMÁNTICA

Las tecnologías previamente expuestas resultan el camino para lograr enlazar e integrar datos en la web, acciones que también se denominan bajo el rótulo de servicio de datos enlazados o *Linked Data* (LD). Las tecnologías de datos enlazados tratan de utilizar la web para crear vínculos entre datos desde distintas fuentes. Desde el punto de vista técnico, los datos enlazados se refieren a la publicación de datos en la web de tal

manera que son legibles por una máquina, ya que tienen definida una semántica explícita, y pueden vincularse o ser vinculados con *data sets* externos.

De esta manera, la integración de tecnologías semánticas a fuentes de datos ya existentes en la web se ha transformado en un campo activo de investigación y aplicación para diversas disciplinas, especialmente en ciencias de la información, hecho que fomenta la publicación e interoperabilidad de datos en la web (Torre-Bastida, González-Rodríguez y Villar-Rodríguez, 2015). Si consideramos adicionalmente que la mayoría de los datos que están publicados en la web se almacenan en bases de datos relacionales, es posible afirmar que la transformación o adaptación de datos de modelos relacionales a RDF es uno de los principales desafíos para el desarrollo de la web semántica, y uno de los más activos campos de investigación en la materia (Hert, Reif y Gall, 2011).

A raíz de lo anterior, se han establecido metodologías y herramientas que proponen la conversión de datos de una base de datos relacional a un modelo basado en una ontología y descrito en RDF. Incluso la W3C ha creado un grupo de investigación denominado RDB2RDF Incubator Group que busca establecer los estándares para traducir bases de datos relacionales a RDF.

En este marco, como se explicita en la Figura 1, se identifican dos mecanismos conceptualmente distintos para efectuar la integración de tecnologías semánticas desde una base de datos relacional: la primera, denominada *transformación*, es utilizada para extraer una ontología a partir de una base de datos relacional y la segunda, denominada *mapping*, efectúa un cruce entre una ontología existente y la base de datos relacional, para establecer vínculos entre las tablas de la base de datos y las clases de la ontología (Hazber et al., 2016).

FIGURA 1.  
Transformación v/s Mapping.



Elaboración propia.

## WEB SEMÁNTICA Y PROPIEDAD INDUSTRIAL

Si el interés por incorporar tecnología semántica o de *linked data* a diversas fuentes de datos ha sido creciente, el caso de la propiedad industrial y las patentes no ha sido diferente.

En primer lugar, actualmente las bases de datos de patentes almacenan la información principalmente en bases de datos relacionales, y exhiben sus datos en formatos CSV o XML, lo que dificulta los posibles análisis semánticos (Bermudez-Edo et al., 2013). Por tanto, es indispensable contar con tecnologías que fortalezcan el intercambio e interoperabilidad de los datos relacionados a la propiedad industrial, debido al potencial de



esta información técnica para los actores involucrados en el ecosistema de la investigación y el desarrollo. En este contexto surge el rol de las tecnologías semánticas en el mundo de las patentes (Diamantini et al., 2013).

Aunque en muchos casos las oficinas de propiedad industrial y otras empresas relacionadas declaran ofrecer servicios de búsqueda semántica de patentes, estas funcionalidades están referidas a sinonimia, variaciones de palabras u otros mecanismos, pero no apuntan a tecnologías basadas en *linked data*. Por otra parte, también existen desarrollos orientados hacia la combinación semántica de *keywords* para fortalecer los análisis de tendencias tecnológicas en patentes o sistemas de análisis semánticos basados en *big data* (An et al., 2018; Shin, Lee y Wang, 2017).

A diferencia de los servicios referidos, un servicio basado en *linked data* resulta prometedor a la hora de soportar la gestión de la I+D+i (García Moreno, 2015). Particularmente, en el entorno de las patentes contribuye a establecer mejores clasificaciones basadas en una ontología, enriquecer los datos con contenido semántico y promover su (re)utilización para entregar ventajas competitivas a las empresas o instituciones que requieran su uso. Sin embargo, el potencial de estas tecnologías se ha tratado fundamentalmente desde el plano de la investigación científica (Angrosh, Cranefield y Stanger, 2014), pero de manera muy escasa en el plano de la integración o aplicación en servicios vigentes.

Dentro de las aplicaciones más concretas de integración de tecnologías de web semántica al mundo de las patentes destaca GoPatents, que es un prototipo orientado a la búsqueda de patentes apoyado por la extracción terminológica, la categorización de patentes y el uso de ontologías de diversos dominios para establecer relaciones (Eisinger, Mönnich y Schroeder, 2014).

Por otro lado, destaca el desarrollo de ciertas ontologías aplicadas al dominio de la I+D+i como GI2MO (Generic Idea and Innovation Management Ontology) basada en RDF y OWL, que incorpora terminología de distintas fuentes y tiene como objetivo vincularse con otros vocabularios estándar como FOAF (Friend of a Friend) o DCTerms, o la ontología IDEA, basada en OWL, que también reutiliza otros vocabularios de uso común para fortalecer la gestión de ideas de innovación (García Moreno, 2015).

Particularmente, en el caso de las patentes resalta el trabajo del grupo de investigación Agile Knowledge Engineering and Semantic Web (AKSW), quienes crearon una ontología para el contexto de la Oficina de Patentes y Marcas de los Estados Unidos (USPTO), la cual denominaron US Patents. Dicha ontología aborda distintas clases y propiedades relacionadas con los documentos de patentes de la USPTO y es utilizada para generar tripletas en RDF que puedan nutrir de contenido semántico a dicha documentación.

En base a lo expuesto, la necesidad por optimizar las bases de datos de patentes es notoria, y las tecnologías semánticas suponen una oportunidad con mucho potencial, a pesar de las características heterogéneas y complejas que pueden presentar estas fuentes de información. Sin embargo, su aplicación no ha sido una práctica masiva y eso da cabida a que sea un campo de investigación con mucho potencial.

## METODOLOGÍA DE INTEGRACIÓN DE TECNOLOGÍAS SEMÁNTICAS

Con base en lo expuesto previamente, se propone establecer una estrategia metodológica de traducción de una base de datos relacional a un modelo semántico basado en un *mapping* utilizando estándares de la W3C como R2RML, que es un lenguaje para expresar asignaciones personalizadas de bases de datos relacionales a conjuntos de datos RDF (Hazber et al., 2016), y una ontología adecuada para el dominio de conocimiento de la propiedad industrial. Dicha estrategia consta de cuatro etapas (Figura 2):

FIGURA 2.  
Estrategia metodológica de integración semántica



Elaboración propia.

En la primera etapa, se propone efectuar un análisis de la base de datos relacional con el objetivo de identificar los campos descriptivos más relevantes, con el propósito de determinar las potenciales relaciones que tendría este modelo de datos con una ontología del mismo dominio de conocimiento. A continuación, se propone la selección de una ontología y, en caso de ser necesario, la adaptación de esta misma para integrar al modelo. Posteriormente, se efectúa el *mapping* o la traducción de la base de datos relacional a un modelo RDF, estableciendo un vínculo entre las tablas de la base de datos y las clases de la ontología, usando Karma, un *software* libre para la integración de datos. Finalmente, se definen y se publican las tripletas resultantes en RDF.

#### ANÁLISIS DE BASE DE DATOS RELACIONAL

En primera instancia, es preciso analizar la base de datos relacional, para lo cual se recomienda revisar el modelo de datos, desde el cual se determinan los elementos relevantes y susceptibles a ser vinculados en un contexto de web semántica. Un modelo de datos permite identificar y describir la estructura de los datos que componen dicha base y la forma en que están relacionados (Figura 3).



PDF generado a partir de XML-JATS4R por Redalyc  
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto



El análisis de la base de datos relacional nos da cuenta de las entidades y actores fundamentales en el ecosistema de la propiedad industrial y los elementos descriptivos o metadatos de cada uno de ellos, que pueden ser extrapolables a un contexto de datos enlazados. Por ejemplo, podemos identificar la entidad “Patente”, que tiene metadatos como título, fecha, número de solicitud, inventor, solicitante, clasificación, entre otros. Asimismo, la entidad “Inventor” tiene un ID, nombre, nacionalidad, entre otros metadatos.

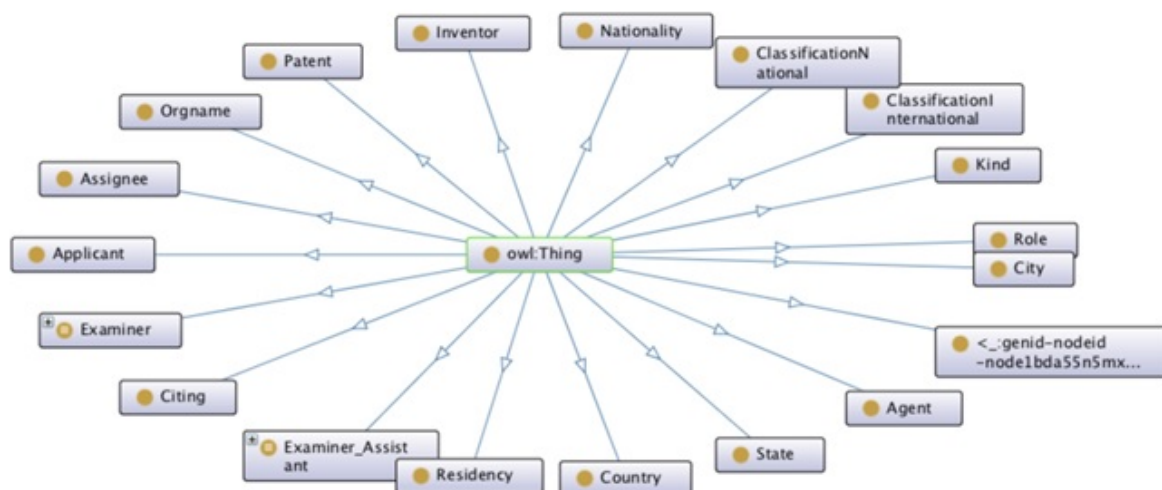
## DEFINICIÓN / ADAPTACIÓN DE ONTOLOGÍA

Para lo anterior, es posible desarrollar una ontología propia o utilizar alguna preexistente para adaptarla al contexto en el cual será utilizada, como la ontología desarrollada por el grupo de investigación AKSW para la USPTO, denominada US Patents Ontology. La naturaleza estandarizada y normalizada de los documentos de patentes permite que la ontología de AKSW u otras similares tengan muchos elementos aplicables a diversos contextos de propiedad industrial.

Para efectuar de manera eficiente el proceso de modelamiento y traducción de la base de datos de patentes a un modelo RDF, es necesario analizar y comprender la estructura de la ontología a utilizar en el procedimiento. Lo anterior puede ser abordado utilizando el *software* de gestión de ontologías *Protégé*. Este software libre tiene como principales funciones la edición, mantención y el análisis descriptivo de una

ontología (Musen, 2015), entregando información relevante para el *mapping*, como sus clases, relaciones, axiomas, entre otros elementos.

FIGURA 4.  
Visualización de clases de la US Patent Ontology en Protégé



Elaboración propia.

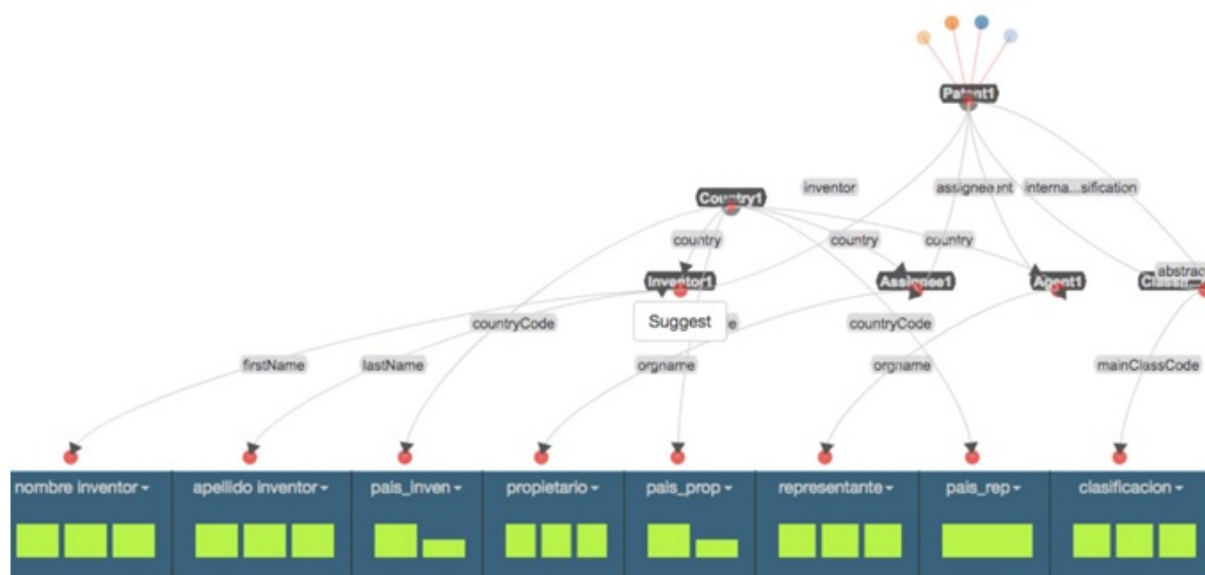
El identificar, analizar y adaptar (en caso de ser necesario) las clases y las propiedades de una ontología resulta fundamental para establecer los parámetros sustanciales que determinarán la traducción o el *mapping* de la base de datos relacional a un modelo RDF, debido a que estos elementos serán interconectados con las entidades de la base de datos original.

## MAPPING / MODELAMIENTO DE DATOS

Posterior al reconocimiento de los elementos de la ontología, es preciso hacer la traducción de la base de datos relacional a un esquema o modelo de datos en RDF. Actualmente existen múltiples procedimientos –más o menos complejos– que permiten publicar datos en RDF directamente desde una base de datos relacional. Por ejemplo, existen herramientas como D2R, que permiten convertir fácilmente la estructura de una base de datos en RDF, aunque sin un fuerte componente de descripción de contenidos semánticos por la carencia de relaciones que ofrece una ontología. Sin embargo, también es posible efectuar la traducción utilizando como estructura base una ontología dada, lo que entrega un mayor potencial semántico. Lo anterior permite generar tripletas en RDF vinculadas a una ontología, que puede ser creada o se puede reutilizar una ya existente, para describir ciertas fuentes de datos de un determinado dominio de conocimiento (Knoblock et al., 2012).

La conversión a un modelo RDF es un paso relevante para publicar fuentes de datos enlazados y los mecanismos antes expuestos permiten hacer la conversión con la semántica subyacente de manera explícita. Por otra parte, existen otros sistemas como R2R o R2RML, que definen lenguajes para traducciones específicas entre distintas fuentes de datos. Dichos lenguajes pueden ser escritos vía códigos o pueden aplicarse de manera semiautomática con alguna herramienta de integración como Karma, un *software* libre que permite transformar o publicar datos en RDF, con base en una ontología, utilizando R2RML a partir de diversas fuentes de datos (bases de datos relacionales, hojas de cálculo, XML, entre otros).

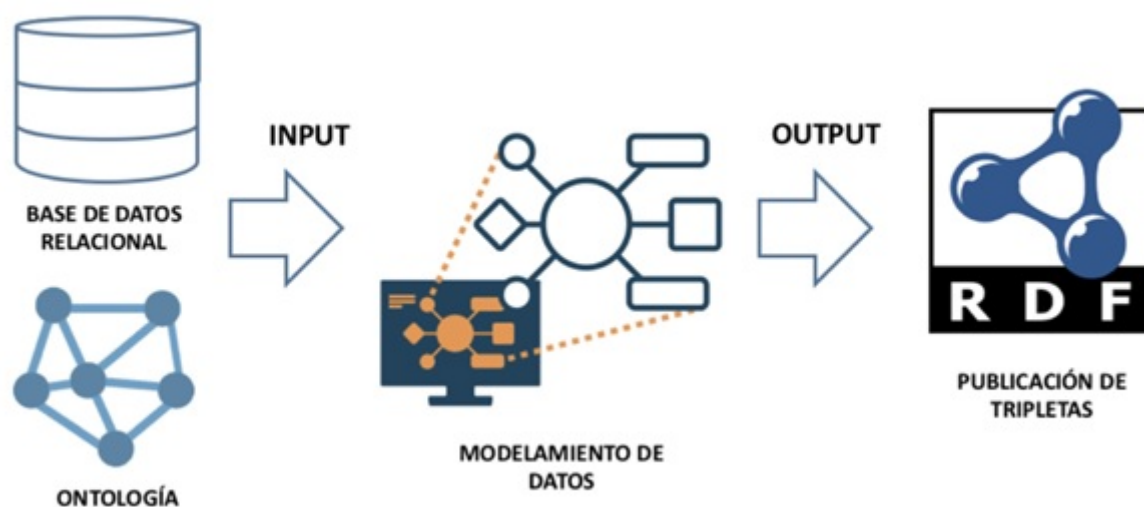
FIGURA 5:  
Construcción de grafo en Karma



Elaboración propia.

Para efectuar el proceso mediante Karma se debe integrar como *input* al *software* la fuente de datos, y una ontología a elección del usuario que permite el modelado a través de una interfaz gráfica que automatiza el proceso, aunque también permite establecer las relaciones entre la base de datos y la ontología de forma manual en una interfaz. Posteriormente, el *software* realiza el mapeo de datos de acuerdo a las clases de la ontología y propone un modelo para vincular estas clases con las tablas del modelo relacional. Luego, Karma construye un grafo que exhibe las posibilidades de mapeo entre la fuente de datos y la ontología para refinar por último el modelo (Figura 5). Finalmente, una vez que el modelo está completo, Karma puede publicar los datos integrados como RDF u otro formato a elección del usuario (Figura 6).

FIGURA 6.  
Modelo de traducción de base de datos a RDF mediante Karma



Elaboración propia.

En términos prácticos, el proceso de modelamiento en Karma se resume en cinco pasos fundamentales: 1) *Input* o integración de ontología y fuente de datos, donde se incorporan los elementos que van a ser parte del modelamiento, 2) Vinculación de clases de la ontología con la fuente de datos, donde se enlazan las entidades de la fuente original con las clases de la ontología, 3) Construcción del grafo, donde se establecen las relaciones semánticas entre las entidades/clases, 4) Refinamiento del modelo, donde se efectúa una revisión y optimización del esquema antes de ser completado y 5) Publicación del modelo, donde resulta un modelo de *mapping* escrito y basado en R2RML. Dicho modelo puede ser conservado y posteriormente replicado.

## DEFINICIÓN DE TRIPLETAS

Finalmente, junto con publicar el modelo R2RML, Karma permite exportar los datos en RDF (además de otros formatos como JSON o Turtle) para generar las tripletas de manera automatizada. Lo anterior es el paso final al proceso de traducción de una base de datos relacional a un modelo de tripletas en RDF, lo que permite nutrir los datos de patentes de un contenido semántico. De esta manera, en el ámbito de la propiedad industrial, se concreta la traducción de un modelo clásico de base de datos relacional (Entidad - atributo - valor) a un modelo en RDF semántico (Sujeto - predicado - objeto) que responde a los estándares de datos enlazados.

Lo anterior permite transformar radicalmente los mecanismos de relación entre los datos, a partir de la aplicación de un modelo de grafo, que favorece los mecanismos de búsqueda y recuperación de información, y permite la integración de datos con otras fuentes.

## CONCLUSIONES

En el contexto actual, un sistema de I+D+i requiere más y mejores herramientas en muchos aspectos: gestión de innovación, de ideas, de información, entre otras. Particularmente, en el caso de la propiedad industrial, los servicios de información tienen muchas oportunidades de fortalecer y mejorar la recuperación de información o la visibilidad de la documentación tecnológica, ya sea protegida o de dominio público, lo que fortalece uno de los objetivos principales de la propiedad industrial, que es incentivar las invenciones a través del acceso a la información.

Una de las grandes oportunidades que se presentan es la inclusión de tecnologías semánticas o de *linked data*, que van tomando cada vez más fuerza y relevancia. Si bien, la mayor parte de los contenidos digitales están publicados en base de datos relacionales, el interés por traducir estas fuentes de información a modelos semánticos ha sido materia creciente de investigación.

En ese marco, la aplicación de mecanismos semiautomáticos de integración de datos resulta un camino válido para que los profesionales de la información y analistas de propiedad industrial exploren en el campo del uso de datos enlazados en el mundo de las patentes, para fortalecer las experiencias de búsqueda y relación semántica de datos en este sector.

## REFERENCIAS

- Adams, S. (2012). *Information Source in Patents*. Berlin: De Gruyter.
- An, J., Kim, K., Mortara, L., y Lee, S. (2018). Deriving technology intelligence from patents: Preposition-based semantic analysis. *Journal of Informetrics*, 12(1), 217-236. <https://doi.org/10.1016/j.joi.2018.01.001>

- Angrosh, M. A., Cranefield, S., y Stanger, N. (2014). Contextual information retrieval in research articles: Semantic publishing tools for the research community. *Semantic Web*, 5(4), 261-293. <https://doi.org/10.3233/SW-130097>
- Bermudez-Edo, M., Noguera, M., Garrido, J. L., y Hurtado, M. V. (2013). Semantic patent information retrieval and management with OWL. *Advances in Intelligent Systems and Computing*, 206 AISC, 33-42. [https://doi.org/10.1007/978-3-642-36981-0\\_4](https://doi.org/10.1007/978-3-642-36981-0_4)
- Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H., y Secret, A. (1994). The World-Wide Web. *Communications of the ACM*, 37(8), 76-82.
- Decker, S., Melnik, S., Van, H., Fensel, D., Klein, M., Broekstra, J., ... Horrocks, I. (2000). Semantic Web: The roles of XML and RDF. *IEEE Internet Computing*, 4(5), 63-74. <https://doi.org/10.1109/4236.877487>
- Diamantini, C., Potena, D., Proietti, M., Smith, F., Storti, E., y Taglino, F. (2013). A semantic framework for knowledge management in virtual innovation factories. *International Journal of Information System Modeling and Design*, 4(4), 70-92. <https://doi.org/10.4018/ijismd.2013100104>
- Eisinger, D., Mönnich, J., y Schroeder, M. (2014). Developing semantic search for the patent domain (Vol. 1292). Presentado en *CEUR Workshop Proceedings*.
- Etzkowitz, H., y Leydesdorff, L. (2000). The dynamics of innovation: From National Systems and “mode 2” to a Triple Helix of university-industry-government relations. *Research Policy*, 29(2), 109-123.
- García Moreno, C. (2015). *Desarrollo de un modelo para la gestión de la I+D+i soportado por tecnologías de la web semántica* (Tesis doctoral). Universidad de Murcia. <http://hdl.handle.net/10803/327316>
- Gerber, A. J., Barnard, A., y Van, D. M. (2007). Towards a semantic web layered architecture (pp. 353–362). Presentado en *Proceedings of the LASTED International Conference on Software Engineering*, SE 2007.
- Hazber, M. A. G., Li, R., Gu, X., y Xu, G. (2016). Integration mapping rules: Transforming relational database to semantic web ontology. *Applied Mathematics and Information Sciences*, 10(3), 881-901. <https://doi.org/10.18576/amis/100307>
- Hert, M., Reif, G., y Gall, H. C. (2011). A comparison of RDB-to-RDF mapping languages (pp. 25-32). Presentado en *ACM International Conference Proceeding Series*. <https://doi.org/10.1145/2063518.2063522>
- Hunt, D., Nguyen, L., y Rodgers, M. (2012). *Patent Searching: Tools & Techniques*. New Jersey: John Wiley & Sons.
- Knoblock, C. A., Szekely, P., Ambite, J. L., Goel, A., Gupta, S., Lerman, K., ... Mallick, P. (2012). Semi-automatically mapping structured sources into the semantic web. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7295 LNCS, 375-390. [https://doi.org/10.1007/978-3-642-30284-8\\_32](https://doi.org/10.1007/978-3-642-30284-8_32)
- Labra Gayo, J. E. (2011). *Web semántica: comprendiendo el cambio hacia la web 3.0*. La Coruña: Netbiblo.
- Musen, M. A. (2015). The Protégé Project: A Look Back and a Look Forward. *AI matters*, 1(4), 4-12. <https://doi.org/10.1145/2757001.2757003>
- OMPI. (2013). Las patentes: Fuente de información tecnológica. Recuperado de [http://www.wipo.int/edocs/pubdocs/es/patents/434/wipo\\_pub\\_1434\\_02.pdf](http://www.wipo.int/edocs/pubdocs/es/patents/434/wipo_pub_1434_02.pdf)
- OMPI. (2016). Principios básicos de la propiedad industrial. Recuperado 4 de marzo de 2018, a partir de [http://www.wipo.int/edocs/pubdocs/es/wipo\\_pub\\_895\\_2016.pdf](http://www.wipo.int/edocs/pubdocs/es/wipo_pub_895_2016.pdf)
- Pastor-Sánchez, J. A. (2011). *Tecnologías de la Web Semántica*. Cataluña: Editorial UOC.
- Shadbolt, N., Hall, W., y Berners-Lee, T. (2006). The semantic web revisited. *IEEE Intelligent Systems*, 21(3), 96-101. <https://doi.org/10.1109/MIS.2006.62>
- Shin, J., Lee, S., y Wang, T. (2017). Semantic Patent Analysis System Based on Big Data. En *2017 IEEE 11th International Conference on Semantic Computing (ICSC)* (pp. 284–285). <https://doi.org/10.1109/ICSC.2017.20>
- Singh, V., Chakraborty, K., y Vincent, L. (2016). Patent database: Their importance in prior art documentation and patent search. *Journal of Intellectual Property Rights*, 21(1), 42-56.

Torre-Bastida, A., González-Rodríguez, M, y Villar-Rodríguez, E. (2015). Datos abiertos enlazados (LOD) y su implantación en bibliotecas: iniciativas y tecnologías. *El profesional de la información*, 24(2), 113-120. <http://dx.doi.org/10.3145/epi.2015.mar.04>