Universidad de las Ciencias Informáticas

Facultad 2

OntoIntegra: Método para la integración de datos almacenados en fuentes heterogéneas

Trabajo final presentado en opción al título de

Máster en Informática Avanzada

Autor: Ing. Leandro Tabares Martín

Tutor: Dr.C. Yanio Hernández Heredia

Co-Tutor: Dr.C. Amed Abel Leiva Mederos

La Habana, noviembre de 2018

Agradecimientos

A Dios, el supremo investigador, por permitirme llegar hasta aquí.

A mi familia, por ser ejemplo de perseverancia y apoyo en todo momento.

A mis tutores, por sembrar en mi la semilla de la investigación y ayudarme a lo largo de este proceso.

Al Grupo de Investigaciones de Web Semántica de la UCI, por ayudarme a crecer y desarrollarme como investigador.

A la UCI, por contribuir a mi formación profesional.

Al programa VLIR/UOS, por ser cantera fértil para el desarrollo de la ciencia.

A la Revolución y, especialmente, a nuestro eterno Comandante en Jefe Fidel Castro Ruz, por enseñarnos que un mundo mejor es posible.

Ing. Leandro Tabares Martín

Dedicatoria

A Nataly, mi niña, por ser mi impulso para el resto de la vida.

A mi esposa, por apoyarme y soportarme cada día.

A mi familia, sin ustedes llegar aquí no hubiese sido posible.

Ing. Leandro Tabares Martín

DECLARACIÓN JURADA DE AUTORÍA Y AGRADECIMIENTOS

Declaro por este medio que yo, Leandro Tabares Martín, con carné de identidad 87110425844, soy el autor principal del trabajo final de maestría OntoIntegra: Método para la integración de datos almacenados en fuentes heterogéneas, desarrollada como parte de la Maestría en Informática Avanzada y que autorizo a la Universidad de las Ciencias Informáticas a hacer uso de la misma en su beneficio, así como los derechos patrimoniales con carácter exclusivo.

Y para que así conste, firmo la presente declaración jurada de autoría a los ____ días del mes de _____ del año _____.

_____

Ing. Leandro Tabares Martín

RESUMEN

La integración de fuentes de datos es el problema de interconectar y acceder a fuentes heterogéneas de datos. En la medida en que las organizaciones han evolucionado este problema se ha convertido en un importante campo de investigación tanto para la academia como para la industria. El proyecto "Las Tecnologías de la Información y la Comunicación apoyando los procesos educativos y la gestión del conocimiento en la educación superior" (ELINF) requiere la integración de fuentes heterogéneas de datos para apoyar el proceso de control de autoridades en los sistemas informáticos que utiliza. La presente investigación tiene como objetivo desarrollar un método con componentes semánticos que permita la integración en una aplicación informática de datos relativos al control de autoridades, almacenados de forma heterogénea en las fuentes de datos utilizadas por el proyecto ELINF. En ella se describe el método desarrollado, a la vez que se valida por medio de un caso de estudio al comparar una aplicación desarrollada sin aplicar el método con una aplicación que instanció el método propuesto en cuanto a su capacidad para integrar fuentes heterogéneas de datos.

Palabras clave: Analíticas de datos, OBDA, OBDI, Ontologías, Semántica.

Data sources integration is about interconnecting and accessing heterogeneous data sources. With the evolution of the organizations, this topic has become an important research field both for academy and industry. The project "Information and Communication Tecnologies supporting the educational process and the knowledge management in higher education" (ELINF) needs to integrate heterogeneous data sources for supporting the authority control process in its software. This work aims to develop a method with semantic components to contribute to the integration of authority control related data, required by the ELINF project, which is stored in heterogeneous data sources into a software. The current research describes the developed method. At the same time, it describes the study case used for validating it, through the comparison of a software application developed without the proposed method and another one instantiating the method regarding to its capacity to integrate heterogeneous data sources.

Keywords: *Data analytics, OBDA, OBDI, Ontologies, Semantics*.

ÍNDICE

Índice de figuras

Índice de tablas

El proyecto “Las Tecnologías de la Información y la Comunicación apoyando los procesos educativos y la gestión del conocimiento en la educación superior” (ELINF) tiene como objetivos (Ciudad-Ricardo y cols., 2017):

- Aumentar la capacidad de las universidades asociadas para diseñar y aplicar de manera apropiada las Tecnologías de la Información y la Comunicación (TIC).
- Mejorar las capacidades de las universidades miembro en la gestión de la información aplicada al aprendizaje y la investigación.
- Desarrollar una plataforma nacional que apoye los servicios educacionales y de información de todas las universidades cubanas.

Los sistemas integrados de gestión bibliotecaria, para la gestión de repositorios digitales, para la gestión de investigaciones y entornos virtuales de aprendizaje forman parte del modelo de referencia de ELINF como se ilustra en la figura 1 y contribuyen a la realización de estos objetivos. Las herramientas informáticas que constituyen la instanciación del modelo ilustrado en la figura 1 gestionan, entre otros elementos, artículos científicos y libros.

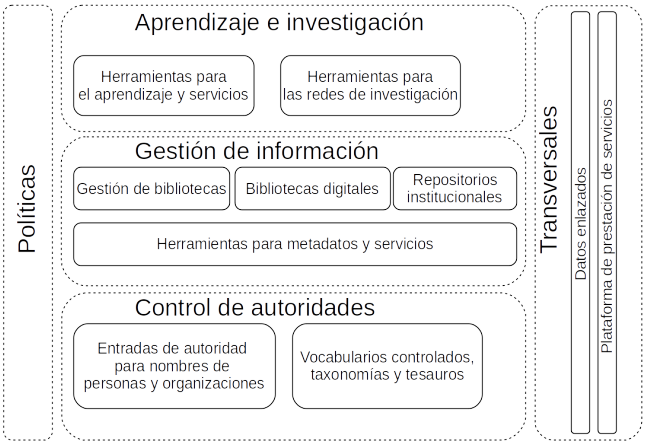

Figura 1: Modelo de referencia del proyecto ELINF. Tomado de (Ciudad-Ricardo y cols., 2017)

En la redacción de artículos científicos se pueden encontrar variaciones de los nombres de un mismo autor, un autor puede tener varios nombres en artículos diferentes y varios autores pueden compartir el mismo nombre. Esta ambigüedad afecta el rendimiento de la recuperación de los documentos, la integración a nivel de bases de datos y puede causar atribuciones de autorías indebidas al aparecer los recursos descritos en los catálogos bajo un autor incorrecto (Han y cols., 2005). En las bibliotecas, museos o en archivos, un catálogo es un conjunto de datos

organizados que describen los recursos de información gestionados por la institución (International Federation of Library Associations and Institutions, 2009), mientras que las personas encargadas de formar estos catálogos se denominan catalogadores (Real Academia Española, 2014). Durante al menos siglo y medio, los catalogadores han documentado sus decisiones acerca de cómo la forma autorizada del nombre de una persona debería ser representada en sus catálogos (Tillett, 2009), sin embargo, según Carrasco y cols. (2016) algunos tipos de inconsistencias que se encuentran a menudo en los nombres registrados en los catálogos son:

- Variantes del mismo nombre.
- Permutaciones del nombre.
- Errores tipográficos.
- Eliminación de palabras vacías.
- Eliminación de diacríticos [1].

Con la finalidad de evitar inconsistencias en los catálogos se realiza el control de autoridades, el cual es uno de los pilares del modelo de referencia del proyecto ELINF (Ciudad-Ricardo y cols., 2017). El control de autoridades es el proceso de seleccionar una forma de un nombre y almacenarla, así como sus variantes y las fuentes de datos utilizadas en el proceso (Sandberg y Jin, 2016). Según Carrasco y cols. (2016) este procedimiento sirve para dos propósitos fundamentales:

- Distinguir creadores que han publicado bajo el mismo nombre por medio de la adición de títulos y otras palabras asociadas con el nombre o incluyendo información sobre las fechas de nacimiento, muerte y actividad del mismo.
- Identificar variantes del nombre o formas alternativas de escribirlo.

En el trabajo de las bibliotecas se ha estado lidiando con la identificación y desambiguación de los nombres de los creadores de los recursos de información desde los comienzos de la catalogación (Harper y Tillett, 2007). Las diferentes formas utilizadas por el mismo creador en varias publicaciones y otros tipos de recursos siempre han acarreado cierto nivel de dificultad para agrupar sus creaciones (Harper y Tillett, 2007).

Los catalogadores han creado registros para el control de autoridades durante décadas, resultando en grandes bases de datos con millones de registros como el “Fichero de Nombres de Autoridades de la Biblioteca del

[1] Dicho de un signo ortográfico: Que sirve para modificar el valor de una letra o de un signo de representación fonética (Real Academia Española, 2014).

Congreso" ("Library of Congress Name Authority File" - LC/NAF, por sus siglas de acuerdo al término en idioma inglés) (Sandberg y Jin, 2016).

Los "Requerimientos Funcionales para Datos de Autoridad" ("Functional Requirements for Authority Data" – FRAD, por sus siglas de acuerdo al término en idioma inglés), elaborados en 2008 por la "Federación Internacional de Asociaciones Bibliotecarias e Instituciones" ("International Federation of Library Associations" - IFLA, por sus siglas de acuerdo al término en idioma inglés), son un modelo entidad-relación enfocado en los datos de autoridad. Los FRAD enmarcan el control de autoridades en términos de entidades y relaciones entre personas y sus nombres; personas y sus obras, manifestaciones, expresiones y elementos (International Federation of Library Associations and Institutions, 2009; Sandberg y Jin, 2016).

El modelo que proponen los FRAD ha sido adoptado por las normas denominadas "Recurso, Descripción y Acceso" ("Resource, Description and Access"- RDA, por sus siglas de acuerdo al término en idioma inglés), que constituyen el código actual de control de autoridades (Sandberg y Jin, 2016).

Iniciativas existentes en la Web contribuyen a proveer mejores mecanismos para identificar a las personas que tienen un rol con respecto a recursos de información (Harper y Tillett, 2007). "Amigo de un amigo" ("Friend of a friend" - FOAF, por sus siglas de acuerdo al término en idioma inglés) es un proyecto que pretende crear una Web de páginas legibles por computadoras describiendo personas, sus vínculos y las cosas que crean y hacen. Encontrar formas de integrar este tipo de iniciativas con los mecanismos existentes en las bibliotecas para el control de autoridades puede contribuir a la inclusión de los catálogos bibliotecarios entre las herramientas disponibles en la Web. Adicionalmente, la disponibilidad de datos bibliotecarios de autoridad en formatos cada vez más compatibles con la Web, tiene el potencial para influenciar de manera positiva la organización de un amplio espectro de contenido disponible en la Web hoy día (Harper y Tillett, 2007).

El "Fichero Internacional Virtual de Autoridades" ("Virtual International Authority File" - VIAF, por sus siglas de acuerdo al término en idioma inglés) combina varios ficheros de nombres de autoridades en un solo servicio de nombres de autoridad. Para lograr esto, VIAF vincula los ficheros de autoridades de varias bibliotecas nacionales y agrupa todos los registros de autoridad para una entidad determinada en un "súper registro de autoridad" que unifica los diferentes nombres de dicha entidad (Online Computer Library Center Inc., 2017).

Otra iniciativa ejecutada con el fin de contribuir al control de autoridades es el "Identificador Internacional Estandarizado de Nombres" ("International Standard Name Identifier" - ISNI, por sus siglas de acuerdo al término en idioma inglés). ISNI es el número estándar global certificado para identificar a los contribuidores de trabajos creativos y a aquellos involucrados en su distribución como investigadores, inventores, escritores, artistas, creadores visuales, actores, productores entre otros (ISNI International Standard Name Identifier, 2017).

La misión de la Autoridad Internacional del ISNI ("ISNI International Authority" - ISNI-IA, por sus siglas de acuerdo al término en idioma inglés) es asignarles a los nombres públicos de autores de recursos de información

un número identificativo persistente, con el objetivo de resolver el problema de la ambigüedad en la búsqueda y recuperación de la información respectiva a nombres de creadores de recursos de información (ISNI International Standard Name Identifier, 2017). Similares a esta iniciativa son las formas utilizadas por el Identificador Abierto de Investigador y Colaborador ("Open Researcher and Contributor ID" - ORCID, por sus siglas de acuerdo al término en idioma inglés) (ORCID, 2017) y el identificador de SCOPUS (Elsevier, 2016).

Con el objetivo de almacenar los datos con vistas a su posterior procesamiento y recuperación, variadas son las estructuras empleadas para realizar de forma eficiente esta tarea (Gutierrez y cols., 2011; Vavliakis y cols., 2013; Lacasta y cols., 2013). Las bases de datos relacionales han sido empleadas durante varias décadas para estructurar los datos y recuperarlos de forma eficiente (Maier, 1983; Shanmugasundaram y cols., 1999; Ilyas y cols., 2004; Spanos y cols., 2012). En los últimos años la utilización de bases de datos "No-SQL" ha ganado popularidad con el objetivo de almacenar datos en forma de documentos (Pokorny, 2013; Moniruzzaman y Hossain, 2013). La utilización del modelo de datos "Marco de trabajo para la descripción de recursos" ("Resource Description Framework" - RDF, por sus siglas de acuerdo al término en idioma inglés) ha contribuido a aportar semántica a los datos almacenados, proveyendo un valor añadido a la información (T. Berners-Lee y cols., 2001; Konstantinou y cols., 2014; Sulé y cols., 2016).

A partir de los análisis realizados por el autor en varias de las principales instituciones bibliotecarias del país, se pudo constatar que en Cuba en el momento en que se realiza la presente investigación, a pesar de las iniciativas registradas por García Rodríguez (2016), no existe un registro central para el control de autoridades, las instituciones realizan este proceso de forma aislada, provocando redundancia en el trabajo. De igual manera, no se reutilizan los registros de autoridades que comparten instituciones extranjeras. Esta situación dificulta la localización de la producción intelectual generada y/o almacenada en la Isla, a la vez que dificulta compartir las entradas de autoridad pertenecientes a autores cubanos con el resto del mundo.

La integración de fuentes de datos es el problema de interconectar y acceder a fuentes heterogéneas de datos (Nachouki y Quafafou, 2011). En la medida en que las organizaciones han evolucionado este problema se ha convertido en un importante campo de investigación tanto para la academia como para la industria (Nachouki y Quafafou, 2011). Aunque la dependencia conceptual es casi universal en el diseño de sistemas de información, también produce una gama de consecuencias negativas incluyendo la inflexibilidad de los sistemas, la heterogeneidad en las formas de almacenar los datos, así como aumentos en los costos de mantenimiento (McGinnes y Kapros, 2015). La heterogeneidad entre dos o más sistemas de bases de datos ocurre cuando estas utilizan diferente infraestructura de software o hardware, siguen diferentes convenciones sintácticas y modelos de representación o cuando interpretan de manera diferente datos similares (Spanos y cols., 2012).

Con el propósito de contribuir a la solución de este problema en el pasado se propuso la creación de una base de datos federada ("Federated database" - FDB, por sus siglas de acuerdo al término en idioma inglés) (Sheth

y Larson, 1990). En arquitecturas de integración de bases de datos típicas, uno o más modelos conceptuales se utilizan para describir los contenidos de cada fuente de datos, las consultas se plantean en concordancia con un esquema conceptual global y para cada fuente de datos, una envoltura ("wrapper", término utilizado en idioma inglés) es responsable de re-formular la consulta y recuperar los datos apropiados (Spanos y cols., 2012; Franke y cols., 2014; El Kadiri y cols., 2015).

Un enfoque para la integración de datos empresariales provenientes de fuentes heterogéneas de datos y descentralizadas es la mediación semántica. Este enfoque por una parte elimina la necesidad de un repositorio central de datos o un esquema federado para todos los datos y, por otra parte, introduce una capa semántica encima de las descripciones de estructuras de datos sintácticas existentes para evitar los conflictos semánticos en la integración (El Kadiri y cols., 2015).

Con el advenimiento de la Web (T. J. Berners-Lee y Cailliau, 1990), la gestión de datos se enfocó en la variedad de información disponible en este nuevo medio. Janev y Vranes (2011) plantearon que las tecnologías semánticas se utilizan en su mayoría para la integración de datos y para mejorar las búsquedas. Según (Hoang y cols., 2014), la integración de datos guiada por ontologías representa una solución flexible, sostenible y extensible. Un paradigma reciente que combina la posibilidad de utilizar razonamiento sobre el conocimiento de un dominio codificado en una ontología, con un mecanismo que permite el uso de la misma ontología para un acceso integrado de alto nivel a las fuentes de datos, es el Acceso e Integración de Datos Basado en Ontologías ("Ontology-Based Data Access and Integration" - OBDA/OBDI, por sus siglas de acuerdo al término en idioma inglés) (Calvanese y cols., 2016, 2017).

Para la realización del control de autoridades a nivel internacional se comparten ficheros utilizando el formato RDF (Sandberg y Jin, 2016), se exponen servicios en la Web (ORCID, 2017), se crean bases de datos locales que luego se exponen por medio de aplicaciones informáticas en la Web (Online Computer Library Center Inc., 2017). La variedad de enfoques utilizados para contribuir a la realización del control de autoridades provoca heterogeneidad estructural entre las fuentes de datos utilizadas en el proceso.

Teniendo en cuenta la heterogeneidad estructural de las fuentes de datos disponibles actualmente para realizar el control de autoridades en el proyecto ELINF se establece como **problema de investigación** el siguiente: ¿Cómo integrar los datos relativos al control de autoridades almacenados de forma heterogénea en las fuentes de datos utilizadas por el proyecto ELINF?

El problema de investigación se enmarca en el **objeto de estudio** la integración de datos almacenados de forma heterogénea y como **campo de acción** el método para integrar datos relativos al control de autoridades en el proyecto ELINF.

Esta investigación se propone como **objetivo general** desarrollar un método con componentes semánticos que permita la integración en una aplicación informática de datos relativos al control de autoridades, almacenados de

forma heterogénea en las fuentes de datos utilizadas por el proyecto ELINF. Como **objetivos específicos** se definen los siguientes:

1. Identificar los referentes teóricos que soportan la integración semántica de datos en aplicaciones informáticas.
2. Desarrollar un método con componentes semánticos que permita integrar en una aplicación informática datos almacenados de forma heterogénea.
3. Validar el método desarrollado mediante un caso de estudio con datos reales relativos al control de autoridades almacenados de forma heterogénea.

Se plantea como **hipótesis de la investigación** que, el desarrollo de un método con componentes semánticos para conducir la integración de datos almacenados de forma heterogénea, contribuye a la integración de los datos relativos al control de autoridades almacenados en las fuentes de datos utilizadas por el proyecto ELINF.

A partir de la hipótesis se identifica como variable independiente el **método con componentes semánticos** el cual se define como el conjunto de actividades (Offermann y cols., 2010) que permitirá conducir el proceso de descripción e integración de datos almacenados de forma heterogénea. Como variable dependiente se identifica la **integración de los datos relativos al control de autoridades** almacenados en las fuentes de datos utilizadas por el proyecto ELINF, definiéndose esta como la capacidad de mostrar en una vista conceptualmente homogénea datos de autoridades almacenados en fuentes heterogéneas de datos.

Variable	Dimensión	Indicadores
Método con componentes semánticos	Etapas que lo componen	Artefactos generados
Integración de los datos relativos al control de autoridades	Fuentes de datos requeridas	Por ciento de las fuentes de datos requeridas que fue posible integrar. Nivel de flexibilidad aportado.

Tabla 1: Operacionalización de las variables de la investigación

Durante toda la investigación se utilizará el método analítico - sintético, el mismo permitirá descomponer las problemáticas que se presenten en sus componentes y relaciones. La síntesis permitirá descubrir las relaciones esenciales existentes entre los componentes de las problemáticas contribuyendo a la sistematización del conocimiento.

El método inductivo - deductivo posee especial relevancia en la formulación de la hipótesis y en la elaboración de conclusiones lógicas. Por medio del método histórico - lógico será posible representar los elementos del estado del arte relevantes a la temática en un orden cronológico que permita comprender la evolución de la misma.

El método hipotético - deductivo de conjunto con el inductivo - deductivo permitirá arribar a conclusiones particulares a partir de la hipótesis, que luego serán validadas a través de un caso de estudio. Esto posibilitará reafirmar la validez de la hipótesis de la investigación.

Por medio de la modelación será posible generar componentes relevantes para la aplicación del método propuesto, estando estrechamente relacionado con el método analítico - sintético y el método sistémico.

El documento de tesis está estructurado en tres capítulos, conclusiones y recomendaciones. A continuación, se brinda una breve descripción de cada uno de los capítulos.

El capítulo 1 aborda los principales referentes teóricos que soportan la integración semántica de datos, en él se refiere la evolución histórica del control de autoridades, se exponen los principales elementos de la Web Semántica y se relatan elementos de la integración de datos que son tomados en consideración para el desarrollo de la investigación.

En el capítulo 2 se identifica el paradigma utilizado en el desarrollo del método propuesto, se realiza la definición del método OntoIntegra a la vez que se describen cada uno de sus componentes. Se detalla el proceso de desarrollo de la ontología creada según la metodología seleccionada, generando cada uno de los artefactos definidos.

El capítulo 3 relata la selección de la estrategia de validación, la preparación del caso de estudio, su diseño con cada una de las etapas que lo componen. Se realiza un análisis de los datos recolectados que permite comprobar la validez de la hipótesis formulada para conducir la investigación.

Capítulo 1

Fundamentación teórica

1.1. Introducción

En este capítulo se abordan los principales referentes teóricos que soportan la integración semántica de datos. En primer lugar se refiere la evolución histórica del control de autoridades con el objetivo de contextualizar el área de aplicación del método que se propone en el presente trabajo. En la sección 1.3 se aborda la Web Semántica describiendo sus principales elementos, los cuales intervienen en la aplicación del método propuesto. Luego se relatan elementos de la integración de datos que se toman en consideración para el desarrollo de la investigación.

1.2. Control de autoridades

El control de autoridades es un problema global que afecta a organizaciones de diversos tipos (Leiva-Mederos y cols., 2013). La comunidad bibliotecaria durante largo tiempo ha sido consciente de la necesidad del control de autoridades (Harper y Tillett, 2007; Tillett, 2009; Leiva-Mederos y cols., 2013; Carrasco y cols., 2016). La necesidad de almacenar de manera uniforme la información correspondiente a cada autor incluido en un catálogo es abordada en el trabajo y la investigación de varias organizaciones internacionales (Leiva-Mederos y cols., 2013). Una panorámica breve del desarrollo en el control de autoridades incluiría los siguientes elementos:

- Se hace explícita la necesidad del control de autoridades y surge la "Cooperación en Nombres de Autoridades" ("Name Authority Cooperative" - NACO, por sus siglas de acuerdo al término en idioma inglés) en la "Biblioteca del Congreso" ("Library of Congress" - LOC, por sus siglas de acuerdo al término en idioma inglés) de los Estados Unidos de América (Leiva-Mederos y cols., 2013). En Asia se establece el "Nombre de Autoridad de Hong Kong Chino" ("Hong Kong Chinese Authority Name" - HKCAN, por sus siglas de acuerdo al término en idioma inglés). Esto significó el reconocimiento de la problemática por dos organizaciones internacionales a partir de los elementos enunciados en el siglo XIX por Cutter (1889).
- Lubetzky y Hayes (1969) mejoran la búsqueda y recuperación de trabajos en catálogos bibliográficos.
- Bregzis (1982) crea el "Número Internacional Estandarizado de Datos de Autoridad" ("International Standard Authority Data Number" - ISADN, por sus siglas de acuerdo al término en idioma inglés) con el fin de vencer las dificultades al recuperar registros bibliográficos con trabajos relativos a un autor en específico y trabajos almacenados bajo un título uniforme.

- La LOC crea el "Fichero de autoridad NACO de la LOC" ("Library of Congress/NACO Authority File" - LC/NAF, por sus siglas de acuerdo al término en idioma inglés) (Sandberg y Jin, 2016).
- La "Federación Internacional de Asociaciones de Bibliotecarios y Bibliotecas" ("International Federation of Library Asociations and Institutions" - IFLA, por sus siglas de acuerdo al término en idioma inglés) crea los FRAD (International Federation of Library Associations and Institutions, 2009).
- El "Centro Bibliotecario Computarizado En Línea" ("Online Computer Library Center" - OCLC, por sus siglas de acuerdo al término en idioma inglés) crea el VIAF (Online Computer Library Center Inc., 2017).

La necesidad de crear registros de autoridad de alta calidad ha impulsado la creación de herramientas como AUTHORIS (Leiva-Mederos y cols., 2013). AUTHORIS aspira a facilitar el procesamiento de datos de autoridad en una manera estandarizada siguiendo los principios de los Datos Enlazados Abiertos (T. Berners-Lee, 2006).

1.3. Web Semántica

La Web 2.0 se basa en un conjunto de tecnologías y estándares definidos por el "Consorcio World Wide Web" ("World Wide Web Consortium" - W3C, por sus siglas de acuerdo al término en idioma inglés) entre los que se encuentra el "Protocolo de Transferencia de Hipertextos" ("Hypertext Transfer Protocol" - HTTP, por sus siglas de acuerdo al término en idioma inglés), los "Identificadores de Recursos Universales" ("Universal Resource Identifier" - URI, por sus siglas de acuerdo al término en idioma inglés) y el "Lenguaje de Marcado de Hipertextos" ("Hypertext Markup Language" - HTML, por sus siglas de acuerdo al término en idioma inglés) para la representación de contenidos (Masinter y cols., 2005). Este último carece de un mecanismo para expresar el significado de los contenidos publicados, imposibilitando a las computadoras procesar los mismos de forma automática (Hidalgo-Delgado, 2015).

El significado del término Web Semántica es definido por T. Berners-Lee y cols. (2001) en un artículo donde expresa: *La Web Semántica no es una Web aparte, sino una extensión de la existente en la que la información posee significado formalmente definido, facilitando que las computadoras y los humanos trabajen cooperativamente*.

1.3.1. Marco de Trabajo para la Descripción de Recursos

Con el objetivo de proporcionar un formato único procesable por computadoras para la representación y descripción de los datos en la Web Semántica, la W3C define el "Marco de Trabajo para la Descripción de Recursos" ("Resource Description Framework" - RDF, por sus siglas de acuerdo al término en idioma inglés), el cual consiste en un modelo de datos simple y totalmente compatible con la Web 2.0 (Motik y cols., 2009; Heath y Bizer, 2011).

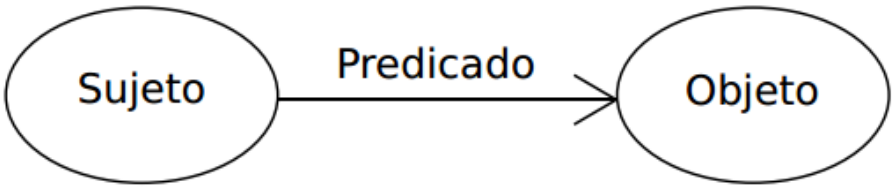

Figura 1.1: Modelo de datos basado en grafo del estándar RDF

Un documento RDF consiste en un conjunto de tripletas de la forma sujeto-predicado-objeto (Figura 1.1) de manera que los datos puedan ser representados en un grafo dirigido donde la primera y la tercera componente corresponde a los nodos del grafo y la segunda componente (predicado) actúa como enlace (arco) entre dichos nodos. Al grafo dirigido descrito con anterioridad se le conoce como Grafo RDF y utiliza las ontologías para la descripción formal de los datos en términos de clases y propiedades (Klyne y Caroll, 2004).

1.3.2. Ontologías

La evolución del RDF en el "Lenguaje de Ontologías Web" ("Web Ontology Language" - OWL, por sus siglas de acuerdo al término en idioma inglés) permite una descripción semántica más rica basada en Lógica Descriptiva (Ian Horrocks y van Harmelen, 2003). El OWL ha sido utilizado en varios escenarios específicos para la construcción de modelos de datos flexibles (H. Agus-Santoso y Abdul-Mehdi, 2011; K. Munir y McClatchey, 2012; Franke y cols., 2014; Sulé y cols., 2016).

El término ontología es utilizado con diferentes significados en diferentes comunidades. Su origen se encuentra en la filosofía y son utilizadas para estudiar la naturaleza del ser y su existencia (Gruber, 1993; Gómez-Pérez y cols., 2004). En computación una definición ampliamente aceptada fue formulada por Gruber (1993) el cual afirma que *una ontología es una especificación explícita de una conceptualización*. Años más tarde Studer y cols. (1998) se basan en este concepto y lo extienden, afirmando que *una ontología es una especificación formal y explícita de una conceptualización compartida*. Según Studer y cols. (1998) una ontología:

- Es explícita porque define los conceptos, propiedades, relaciones, funciones, axiomas y restricciones que la componen.
- Es formal porque es legible e interpretable por computadoras.
- Es una conceptualización porque es un modelo abstracto y una vista simplificada de los elementos reales que representa.
- Es compartida porque se ha arribado previamente a un consenso sobre la información y es aceptada por un grupo de expertos.

Staab y Studer (2009) afirman que existen cuatro tipos diferentes de ontologías a diferentes niveles de granularidad. En un nivel superior se encuentran las ontologías fundacionales, las cuales capturan conceptos generales independientes de un dominio específico. En un segundo nivel de abstracción se encuentran las ontologías de dominio, las cuales modelan conceptos y relaciones que son relevantes para un dominio específico. En estas ontologías se suelen utilizar los términos de la ontologías fundacionales.

Otro tipo de ontología son las ontologías de tareas, las cuales describen conceptos de una tarea en específico. A un nivel más bajo de abstracción se encuentran las ontologías de aplicaciones. Estas combinan ontologías de dominio y ontologías de tareas extendiéndolas con nuevos conceptos y relaciones más específicos (Staab y Studer, 2009).

Gruber (1993) y Gómez-Pérez y cols. (2004) proponen que las ontologías se modelen utilizando cinco tipos de componentes: clases, relaciones, funciones, instancias y axiomas formales. Las **clases** representan conceptos que pueden ser abstractos o no. Las clases de una ontología comúnmente se organizan en taxonomías a través de las cuales se pueden aplicar mecanismos de herencia.

Las **relaciones** representan tipos de asociaciones entre conceptos del dominio. Formalmente se pueden definir como cualquier subconjunto del producto de n conjuntos, esto es $R \subset C_1 \times C_2 \times \ldots \times C_n$. Las ontologías con frecuencia poseen relaciones binarias, donde el primer argumento es conocido como el dominio y el segundo como el rango (Gómez-Pérez y cols., 2004).

Las **funciones** son un caso especial de relaciones donde el n-ésimo elemento de la relación es único para los $n-1$ elementos precedentes, es decir, $F : C_1 \times C_2 \times \ldots \times C_{n-1} \rightarrow C_n$ (Gómez-Pérez y cols., 2004).

Las **instancias** se utilizan para representar los individuos en una ontología (Gómez-Pérez y cols., 2004).

Según Gruber (1993) los axiomas formales sirven para modelar sentencias que siempre son verdaderas. Normalmente se usan para representar conocimiento que no puede ser definido formalmente por otros componentes. Además, los axiomas formales se usan para verificar la consistencia de la ontología en sí misma o la consistencia del conocimiento almacenado en una base de conocimiento (Gómez-Pérez y cols., 2004).

1.3.3. Metodologías para el desarrollo de ontologías

La Ingeniería Ontológica es la disciplina que estudia los principios, métodos y herramientas para crear y mantener ontologías. Una metodología de Ingeniería Ontológica proporciona el aspecto metodológico del desarrollo de ontologías. Con el objetivo de asistir a los ingenieros ontológicos y a los expertos de dominio en la construcción de ontologías se han desarrollado varias metodologías para la Ingeniería Ontológica (Iqbal y cols., 2013).

Un análisis en profundidad de las principales metodologías para el desarrollo de ontologías es realizado por Iqbal y cols. (2013), estos autores proponen ocho criterios para la comparación de metodologías que se detallan a continuación.

Criterio 1. Tipo de desarrollo

La literatura revela que las metodologías para el desarrollo de ontologías pueden dividirse en tres categorías fundamentales: modelos basados en etapas, modelos de prototipos evolutivos y guías; dependiendo del tipo de modelo de desarrollo que siguen. Los diferentes enfoques tienen sus aspectos positivos y negativos. Las metodologías basadas en etapas pueden ser factibles en escenarios donde el propósito y los requerimientos están claramente definidos. Por el contrario, los prototipos evolutivos pueden ser la mejor elección cuando los requerimientos no están claramente definidos desde el inicio y es necesario refinar la ontología con el paso del tiempo. Las guías mayormente se enfocan en hacer sugerencias útiles, recomendar reglas y técnicas con el objetivo de tomar mejores decisiones en lugar de enfocarse en el modelo de desarrollo en sí.

Criterio 2. Soporte para la construcción colaborativa

Las ontologías pueden construirse tanto de forma aislada como colaborativamente. El soporte para la construcción colaborativa permite a diferentes miembros del equipo de desarrollo de ontologías trabajar en una misma ontología a la vez. Los miembros del equipo de desarrollo pueden estar dispersos geográficamente sin afectar la eficiencia del proyecto.

Criterio 3. Soporte para re-utilización

El desarrollo de ontologías es una tarea costosa en tiempo y esfuerzo. Con el fin de economizar estos factores la noción de ontologías re-utilizables ha cobrado fuerzas a lo largo del tiempo. Las metodologías que soportan la re-utilización les permiten a los equipos de desarrollo de ontologías hacer uso de ontoloogías ya existentes reduciendo el tiempo y esfuerzo empleado en el desarrollo. El ahorro del tiempo les permite a los ingenieros enfocarse en los defectos presentes en las ontologías existentes, mejorando su calidad.

Criterio 4. Soporte para interoperabilidad

Las ontologías de dominio que se desarrollen utilizando metodologías que soporten la interoperabilidad, proveerán a los sistemas que las usen de los mismos conceptos de alto nivel, por lo que les será más fácil comunicarse y compartir la información que gestionan entre sí.

Critero 5. Grado de dependencia de la aplicación

Diferentes metodologías durante el proceso de desarrollo pueden adoptar diferentes enfoques en cuanto a la dependencia de una aplicación. Una metodología puede optar por uno de tres escenarios: dependiente de la aplicación (la ontología se desarrolla enfocada en la base de conocimiento para una aplicación específica), semi-independiente de la aplicación (la ontología se desarrolla teniendo en cuenta posibles escenarios de

aplicación durante la etapa de especificación) e independiente de la aplicación (la ontología se desarrolla sin enfocarse en un sistema en particular).

Criterio 6. Recomendación del ciclo de vida

El ciclo de vida de una ontología identifica el conjunto de etapas por las que pasa una ontología durante su vida. Varias metodologías no recomiendan claramente un ciclo de vida.

Criterio 7. Estrategias para la identificación de conceptos

La identificación de los conceptos candidatos para la inclusión en la ontología es indudablemente un proceso crucial en el desarrollo de la misma. Existen técnicas para la identificación de conceptos, algunas de ellas utilizan un enfoque abajo - arriba, mientras que otras emplean un enfoque arriba - abajo y unas terceras el enfoque que siguen es del centro hacia afuera.

Criterio 8. Nivel de detalle de la metodología

Cada metodología comprende algunas actividades y técnicas para soportar el desarrollo de ontologías. El análisis de la literatura ha revelado la existencia de metodologías que no proveen suficientes detalles para el empleo de sus actividades y técnicas. Para propósitos de análisis, este trabajo clasificará las metodologías en tres grupos de acuerdo a este criterio: suficientes detalles, algunos detalles e insuficientes detalles. Las metodologías que no brindan detalles o estos son muy vagos se clasificarán como de insuficientes detalles. Por otra parte, las metodologías que no cubren completamente los detalles pero al menos proporcionan algunos detalles sobre sus actividades y técnicas se clasificarán como algunos detalles. Asimismo, las metodologías clasificadas como suficientes detalles proveen un razonable nivel de detalles sobre las actividades y técnicas que emplean, permitiendo al lector comprenderlas claramente.

Una clasificación de las metodologías analizadas de acuerdo a los criterios anteriormente expuestos se muestra en la tabla 1.1. Posterior al análisis de diferentes metodologías para el desarrollo de ontologías se decidió utilizar para el presente trabajo METHONTOLOGY ya que recomienda un ciclo de vida, es reutilizable y provee suficientes detalles sobre las técnicas y actividades empleadas en ella.

Metodologías	Tipo de Desarrollo	Construcción colaborativa	¿Reutilizable?	Dependiente de la aplicación	Ciclo de vida	Estrategias para identificar los conceptos	Nivel de detalle	¿Interoperable?
TOVE	Basada en etapas	No	Sí	Semi independiente	No	Media	Algunos detalles	No
Enterprise model approach	Basada en etapas	No	Sí	Independiente	No	Media	Algunos detalles	No
METHONTOLOGY	Prototipo evolutivo	No	Sí	Independiente	Sí	Media	Suficientes detalles	No
KBSI IDEF5	Prototipo evolutivo	No	Sí	Independiente	No	No es clara	Algunos detalles	No
Ontolingua	Desarrollo modular	Sí	Sí	Independiente	No	No es clara	Algunos detalles	Sí
KACTUS	Desarrollo modular	No	Sí	Dependiente	No	Estrategia arriba-abajo	Insuficientes detalles	No
PLINIUS	Basada en guías	No	No	Independiente	No	Estrategia abajo-arriba	Algunos detalles	No
ONIONS	Desarrollo modular basado en guías	No	No	Dependiente	No	No es clara	Insuficientes detalles	Sí
Mikrokosmos	Basada en guías	No	No	Dependiente	No	Estrategia basada en reglas	Algunos detalles	No
MENELAS	Basada en guías	No	No	Dependiente	No	Grafos de conceptos	Insuficientes detalles	No
SENSUS	No menciona preferencias	Sí	Sí	Semi independiente	No	Abajo-arriba	Algunos detalles	Sí
Cyc	Prototipo evolutivo	No	Sí	Independiente	No	No es clara	Algunos detalles	No
UPON	Prototipo evolutivo	No	Sí	Independiente	Sí	Media	Algunos detalles	No
Método 101	Prototipo evolutivo	No	Sí	Independiente	No	Consenso del desarrollador	Algunos detalles	No
On-To-Knowledge	Prototipo evolutivo	No	No	Dependiente	Sí	Media	Algunos detalles	No

Tabla 1.1: Comparación de las metodologías analizadas. Tomado de (Iqbal y cols., 2013)

1.3.4. Datos Enlazados Abiertos

Los Datos Enlazados Abiertos son un conjunto de principios y buenas prácticas para la publicación de datos en la Web. Estos datos pueden estar dispersos geográficamente y pertenecer a una o varias organizaciones. En este sentido T. Berners-Lee (2006) propone cuatro principios:

1. Utilizar URIs como nombres para las cosas.
2. Utilizar URIs HTTP para que las personas puedan buscar esos nombres.
3. Cuando alguien busca una URI, proveer información útil por medio de los estándares.
4. Incluir vínculos a otras URIs para que se puedan descubrir más cosas.

El primer principio propone el uso de URIs para identificar no solo documentos web y contenido digital, sino que sirva además para referenciar a objetos del mundo real y conceptos abstractos. El segundo principio propone el uso de URIs basadas en el protocolo HTTP para identificar objetos y conceptos abstractos, posibilitando que estas URIs estén desreferenciadas sobre dicho protocolo y en cambio proporcionen una descripción del objeto o concepto identificado. El tercer principio propone el uso del modelo de datos RDF para publicar datos estructurados en la Web. El cuarto principio propone el uso de enlaces para enlazar no solo documentos web, sino cualquier tipo de recurso (Hidalgo-Delgado, 2015).

1.4. Integración de datos

Actualmente la existencia, competitividad y rentabilidad de diversas compañías depende de los flujos de datos. Sin embargo, la variedad de fuentes, tipos y volúmenes de datos ha vuelto más complejo el proceso de encontrarlos (Alooma Inc., 2017). La integración de datos se refiere a las técnicas involucradas en combinar los datos almacenados en diferentes ubicaciones en una vista común integrada (Michel, 2017). La disponibilidad de recursos de información heterogéneos y distribuidos ha conducido a considerar aproximaciones para la integración de datos en los que fuentes de datos independientes puedan participar en federaciones virtuales de datos. Mientras que los almacenes de datos son repositorios rígidos controlados por las premisas de una sola compañía, las nuevas necesidades de información deben acomodarse a la adición oportuna de nuevas fuentes de datos provenientes de instituciones independientes.

Por otra parte, la semántica de los datos almacenados en ocasiones no es descrita por los esquemas de bases de datos. Hasta cierto punto la semántica implícita de los datos se puede inferir a partir de las restricciones de integridad o patrones de diseño de bases de datos comunes, pero la semántica adicional se encuentra codificada en

las aplicaciones que consumen las fuentes de datos. Además, frecuentemente los esquemas de bases de datos se optimizan por cuestiones de rendimiento, resultando en una mezcla de la semántica de los datos con aspectos técnicos. Por estas razones, los métodos utilizados para la integración de datos en la Web deben poseer la capacidad de capturar y compartir conceptualizaciones formales comunes en una manera explícita procesable por computadoras. Esto convencionalmente se logra por medio de la utilización de vocabularios controlados, tesauros y ontologías.

1.4.1. Principios para la integración de datos

Convencionalmente, un sistema para la integración de datos Ψ se denota por la tupla $< \Gamma, \Upsilon, \Sigma >$ donde:

- Γ es el esquema global utilizado para representar la vista unificada.
- Υ son las fuentes de datos representadas por el conjunto de esquemas locales $\upsilon_1, \cdots, \upsilon_n$.
- Σ especifica las correspondencias entre los conceptos de los esquemas locales y los conceptos del esquema global.

Dependiendo de los lenguajes de modelado utilizados para definir los esquemas locales y globales, los conceptos pueden ser clases de una ontología, tablas de una base de datos, objetos, entre otros.

Un sistema para la integración de datos responde a consultas formuladas en términos del esquema global Γ reformulándolas en las consultas correspondientes sobre las fuentes de datos $\upsilon_1, \cdots, \upsilon_n$ por medio de la información almacenada en Σ. Con respecto a la forma en que se expresan los mapeos se han propuesto dos enfoques principales: el enfoque “Global-como-Vista” (“Global-as-View” - GAV, por sus siglas de acuerdo al término en idioma inglés), en el que el esquema global se expresa en términos de consultas (o vistas) sobre los esquemas locales; mientras que en el enfoque “Local-como-Vista” (“Local-as-View” - LAV, por sus siglas de acuerdo al término en idioma inglés), los esquemas globales se expresan en términos de consultas (o vistas) sobre el esquema global. Estos enfoques han sido abundantemente descritos en la literatura (Doan y cols., 2012; Lenzerini, 2002).

Variados son los sistemas propuestos para la integración de datos basados en RDF, unos en forma de motores de consultas federados sobre SPARQL (Schwarte y cols., 2011; Görlitz y Staab, 2010; Corby y cols., 2012; Macina y cols., 2016) y otros como fragmentos de Datos Enlazados (Verborgh y cols., 2016). Sin embargo, los enfoques de estos sistemas no se pueden clasificar claramente como GAV o LAV porque sus objetivos son producir planes de consultas eficientes sobre fuentes de datos distribuidas que soportan SPARQL sin la mediación de un esquema.

1.4.2. Acceso e Integración de Datos Basado en Ontologías

El "Acceso a Datos Basado en Ontologías" ("Ontology-based Data Access" - OBDA, por sus siglas de acuerdo al término en idioma inglés) es un dominio de la Integración de Datos que propone la utilización de ontologías para crear una capa conceptual formal; en esta capa la ontología representa el dominio de los datos almacenados en la fuente de datos y el mapeo describe las relaciones entre la ontología y la fuente de datos (Calvanese y cols., 2015; Kharlamov y cols., 2017). Tabares-Martín, Fernández-Peña, y Leiva-Mederos (2016) definen formalmente un sistema OBDA como una tupla $\Omega =< \tau, \sigma, \mu >$ donde:

- τ es la parte terminológica de la ontología. Se consideran ontologías basadas en Lógica Descriptiva, por lo que τ es una TBox según la Lógica Descriptiva.
- σ es el conjunto de fuentes de datos.
- μ es un conjunto de mapeos, cada uno de la forma $\Phi(x) \leftarrow \Xi(x)$ donde:
  - $\Phi(x)$ es una consulta sobre σ, retornando las tuplas de valor para x.
  - $\Xi(x)$ es una consulta sobre τ cuyas variables libres provienen de x.

La ontología en el paradigma OBDA provee un punto de acceso único para un acceso a datos semántico destinado a los consumidores de datos, a la vez que permite exportar los datos de las fuentes integradas en un formato semántico o ejecutar consultas en términos de un modelo conceptual orientado al usuario que lo abstraen de los detalles a nivel de implementación que comúnmente se encuentran en los esquemas de bases de datos (Kharlamov y cols., 2017). Por otra parte, los expertos en el dominio son capaces de expresar las necesidades de información en sus propios términos sin conocimiento previo sobre la forma en que están estructurados los datos en la fuente, a la vez que reciben las respuestas a sus consultas en los términos definidos en la ontología.

La "Integración de Datos Basada en Ontologías" ("Ontology-based Data Integration" - OBDI, por sus siglas de acuerdo al término en idioma inglés) es un caso más general de OBDA en el cual las organizaciones gestionan de forma integrada diferentes fuentes de datos por medio del paradigma OBDA. La OBDI se ha utilizado en trabajos como los descritos por Calvanese y cols. (2016); Daraio y cols. (2016); Kharlamov y cols. (2016)

1.5. Conclusiones del capítulo

El objetivo que persigue este trabajo es desarrollar un método con componentes semánticos que permita la integración en una aplicación informática de datos relativos al control de autoridades almacenados de forma heterogénea. En este capítulo se abordó el control de autoridades realizando un análisis de su desarrollo que

permitió conocer su evolución e identificar los problemas actuales existentes en la temática. Posteriormente se describieron diferentes elementos que forman parte de las tecnologías de la Web Semántica y que posibilitan describir semánticamente los datos almacenados en fuentes estructuralmente heterogéneas. Luego se examinaron principios para la integración de datos, haciendo énfasis en la integración semántica de datos.

Al realizar la revisión bibliográfica respectiva al estado del arte de la temática se evidenciaron diferentes aproximaciones que posibilitan la integración semántica de datos, siendo la Integración de Datos Basada en Ontologías una de las más promisorias. Sin embargo, a pesar de que se describen los elementos que intervienen en la OBDI, en la literatura revisada no se especifica un método para llevarla a cabo.

Capítulo 2

Método para la integración de datos basada en ontologías

2.1. Paradigma utilizado en el desarrollo del método

La investigación en la disciplina de sistemas de información (SI) se caracteriza por dos paradigmas: las ciencias del comportamiento y las ciencias del diseño (Hevner y cols., 2004). El primer paradigma persigue el desarrollo y verificación de teorías que expliquen o pronostiquen el comportamiento humano u organizacional. El paradigma de las ciencias del diseño tiene como fin la creación de innovaciones que definan ideas prácticas, capacidades tecnológicas y productos a través de los cuales puede lograrse el análisis, diseño, implementación, gestión y uso de sistemas de información de manera efectiva y eficiente (Denning, 1997). Un SI es un sistema basado en computadoras que ayuda a la gestión y uso de la información en una organización o entre varias organizaciones (García Noguera, 2009). Analizando la naturaleza del problema tratado en esta investigación y la relación entre su campo de acción (el método para integrar datos relativos al control de autoridades en el proyecto ELINF) y la disciplina de SI, el desarrollo de la solución se concibió y ejecutó bajo el paradigma de las ciencias del diseño.

Las ciencias del diseño crean y evalúan artefactos orientados a mejorar y entender el comportamiento de los sistemas de información, estos artefactos pueden ser constructos, modelos, métodos e instanciaciones (March y Smith, 1995). Los constructos pertenecen al vocabulario conceptual de un dominio y son empleados por los modelos para representar una situación del mundo real en términos del diseño de un problema y su espacio de solución (Simon, 1996).

Los modelos son abstracciones y representaciones de la realidad (Hevner y cols., 2004). Estos contribuyen a la comprensión de los problemas y las soluciones y frecuentemente representan el vínculo entre el problema y los componentes de la solución permitiendo la exploración de los efectos causados por las decisiones del diseño en el mundo real (Hevner y cols., 2004).

La búsqueda dentro de ese espacio de solución es guiada por métodos, los cuales definen procesos, proveen una guía sobre como resolver problemas. Estos pueden ser algoritmos matemáticos que definen explícitamente el proceso de búsqueda de la solución, descripciones textuales informales de buenas prácticas o combinaciones de ambas (Hevner y cols., 2004).

Las instanciaciones muestran cómo los constructos, modelos o métodos pueden implementarse en un sistema. Ellas demuestran la viabilidad de implementar los métodos y modelos, a la vez que facilitan la evaluación concreta

del artefacto que representan. Por otra parte le permiten a los investigadores aprender sobre el mundo real y cómo el artefacto lo afecta (Hevner y cols., 2004).

Es importante mencionar que para el desarrollo de esta investigación se consideraron además las guías para la investigación en ciencias del diseño, propuestas por Hevner y cols. (2004). Estas guías, fueron establecidas con el propósito de auxiliar a investigadores, revisores, editores y lectores en la comprensión de los requerimientos para una investigación efectiva en ciencias del diseño y se sintetizan el la tabla 2.1.

Guía	Descripción
Guía 1: El diseño como un artefacto	La investigación en ciencias del diseño debe producir un artefacto viable en la forma de un constructo, un modelo, un método o una instanciación.
Guía 2: Relevancia del problema	El objetivo de la investigación en ciencias del diseño es desarrollar soluciones basadas en la tecnología para problemas de negocio importantes y relevantes.
Guía 3: Evaluación del diseño	La utilidad, calidad y eficacia de un artefacto de diseño debe ser rigurosamente demostrada a través de métodos de evaluación bien ejecutados.
Guía 4: Contribuciones de la investigación	La investigación efectiva en ciencias del diseño debe proveer contribuciones claras y verificables en las áreas del artefacto de diseño, fundamentos del diseño y/o metodologías del diseño.
Guía 5: Rigor de la investigación	La investigación en ciencias del diseño se basa en la aplicación de métodos rigurosos tanto en la construcción como en la evaluación del artefacto de diseño.
Guía 6: Diseño como proceso de búsqueda	La búsqueda de un artefacto efectivo requiere la utilización de los métodos disponibles para alcanzar el fin requerido mientras se satisfacen las reglas en el entorno del problema.
Guía 7: Comunicación de la investigación	La investigación en ciencias del diseño debe ser presentada efectivamente tanto a audiencias especializadas como a administrativos.

Tabla 2.1: Guías para la investigación en ciencias del diseño. Tomado de (Hevner y cols., 2004).

La presente investigación fue desarrollada según el proceso definido por Peffers y cols. (2006) para las investigaciones en ciencias del diseño que se ilustra en la figura 2.1. Este proceso propone seis actividades secuenciales: identificación del problema y motivación, objetivos de la solución, diseño y desarrollo, demostración, evaluación y comunicación. La *identificación del problema y motivación* pretende definir el problema de investigación específico y justificar el valor de una solución. Justificar el valor de la solución tiene dos objetivos: motivar al investigador y a la audiencia de la investigación a buscar una solución y contribuir a la comprensión del razonamiento realizado por el investigador sobre el problema investigado. Los recursos requeridos para esta actividad incluyen el conocimiento del estado del problema y la importancia de su solución (Peffers y cols., 2006).

La actividad *objetivos de la solución* tiene como propósito fundamental inferir los objetivos de una solución a partir de la definición de un problema. Los objetivos pueden ser cuantitativos o cualitativos y deben ser inferidos racionalmente a partir de la especificación del problema. Los recursos requeridos para esta actividad pueden incluir el conocimiento del estado actual de problemas, sus soluciones y eficacia de estas últimas (Peffers y cols., 2006).

El *diseño y desarrollo* persigue la creación de la solución por medio de la creación de los artefactos correspondientes. Estos artefactos son, potencialmente, constructos, modelos, métodos o instanciaciones. Esta actividad incluye determinar las funcionalidades necesarias del artefacto y su arquitectura, para posteriormente crear el artefacto. Los recursos requeridos para transitar de los objetivos al diseño y desarrollo incluyen el conocimiento de una teoría que puede convertirse en solución (Peffers y cols., 2006).

La *demostración* pretende probar la eficacia del artefacto para resolver el problema. Esta actividad puede incluir el uso del artefacto en experimentaciones, simulaciones, casos de estudio, pruebas u otras actividades apropiadas. Los recursos requeridos para la demostración incluyen el conocimiento sobre cómo usar el artefacto para solucionar el problema (Peffers y cols., 2006).

En la actividad de *evaluación* se observa y mide cuánto influye el artefacto en la solución del problema. Esta actividad incluye la comparación de los resultados aplicando aproximaciones existentes con los obtenidos al aplicar el artefacto desarrollado. Son requeridos conocimientos sobre métricas relevantes y técnicas de análisis (Peffers y cols., 2006).

La actividad *comunicación* pretende poner en conocimiento de investigadores y otras audiencias relevantes, el problema y su importancia, el artefacto, su utilidad y novedad, el rigor de su diseño y su efectividad (Peffers y cols., 2006).

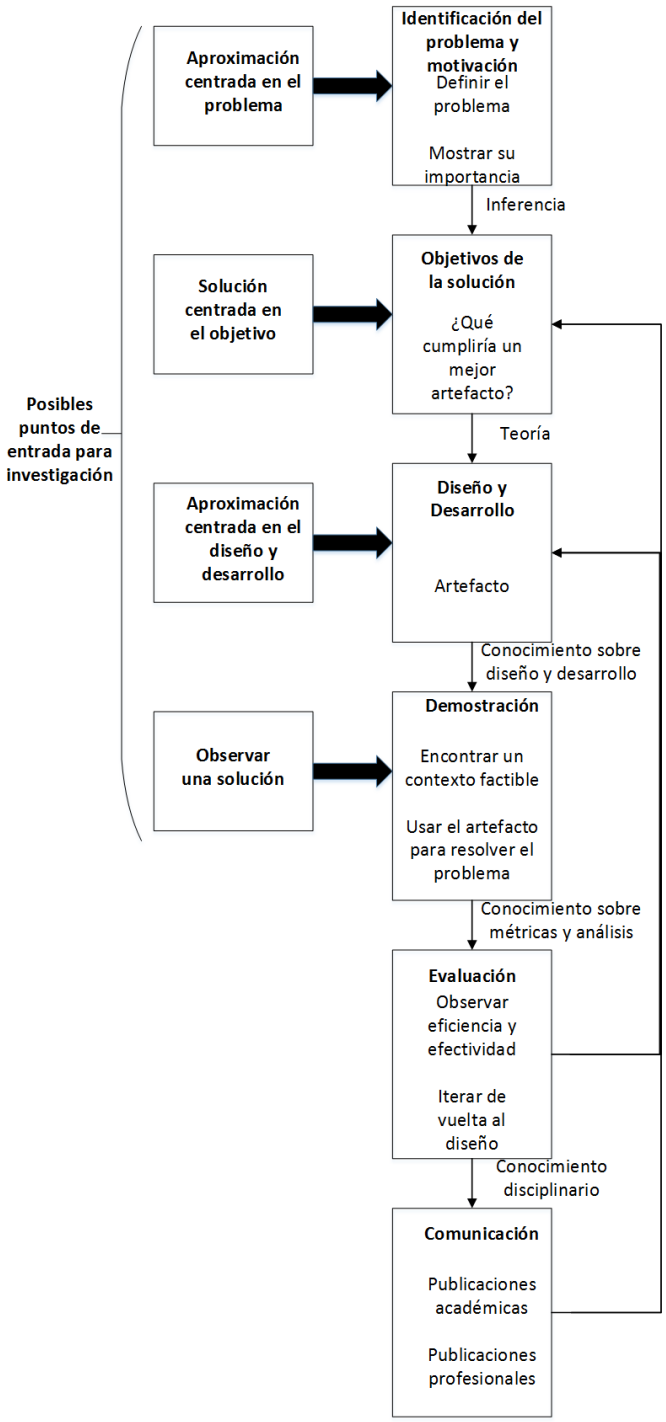

Figura 2.1: Modelo del proceso de investigación en ciencias del diseño. Tomado de (Peffers y cols., 2006)

Los resultados de la primera y segunda actividad fueron expresados en la introducción de este documento. El diseño y desarrollo incluye determinar las funcionalidades deseadas del artefacto y su arquitectura para luego crear el artefacto en cuestión. Los elementos esenciales que debe poseer una aplicación informática para el

OBDA/OBDI fueron sintetizados en el subepígrafe 1.4.2. Los artefactos creados serán descritos en los epígrafes siguientes, mientras que la demostración de su eficacia en la solución del problema, así como la observación y medición de su desempeño relativos a las actividades cuatro y cinco del proceso serán expuestas en el Capítulo 3. Como evidencia de la comunicación a otros investigadores del problema abordado en la investigación y su relevancia, el artefacto creado, su utilidad y su efectividad, se encuentran artículos publicados en revistas y trabajos presentados en eventos por el autor de esta investigación (Tabares-Martín y cols., 2015; Tabares-Martín, Fernández-Peña, y Leiva-Mederos, 2016; Tabares-Martín, Fernández-Peña, Leiva-Mederos, y Nummenmaa, 2016).

El método propuesto para la Integración de Datos Basada en Ontologías se denomina OntoIntegra. Este método describe el proceso para la integración de los datos almacenados en fuentes de datos estructuralmente heterogéneas en una respuesta conceptualmente homogénea, a partir de la instanciación de una ontología creada para la OBDI. El método se basa en los constructos: consulta sintáctica e instancia de una ontología.

Definición 2.1.1 *Una consulta sintáctica es la secuencia textual de instrucciones que permite recuperar información almacenada en al menos una fuente de datos.*

Definición 2.1.2 *Sea una ontología compuesta por una parte conceptual (T-Box) y una parte asercional (A-Box). Se denomina instancia de la ontología a su parte asercional.*

2.2. OntoIntegra

El aporte teórico fundamental de esta investigación es un método para la integración semántica de datos almacenados en fuentes estructuralmente heterogéneas. La figura 2.2 ilustra el método propuesto, denominado OntoIntegra. Este método describe la obtención de información conceptualmente integrada a través de cinco pasos fundamentales:

1. Análisis estructural de cada fuente de datos.
2. Análisis semántico de la información almacenada en cada fuente de datos.
3. Instanciación de la ontología para OBDI.
4. Publicación de la instancia de la ontología creada.
5. Consumo por una aplicación informática de la instancia de la ontología creada.

Figura 2.2: Esquema representativo del método OntoIntegra

2.2.1. Análisis estructural de cada fuente de datos

El análisis estructural de cada fuente de datos tiene como objetivo identificar la forma en que se encuentran estructurados los datos a recuperar, formulando la consulta sintáctica adecuada a este proceso. En este paso interviene un Experto del dominio, el cual es una persona que conoce en detalle la estructura de los datos gestionados por la fuente de datos en cuestión. En el caso de las bases de datos relacionales, el "Lenguaje de Consultas Estructurado" ("Structured Query Language" - SQL, por sus siglas de acuerdo al término en idioma inglés) constituye el estándar para la recuperación de datos (Eisenberg y Melton, 1999). Este tipo de bases de datos se ha utilizado durante más de dos décadas (Buckles y Petry, 1993; Hristidis y Papakonstantinou, 2002; Alfred y cols., 2018), por lo que gran parte de los datos almacenados a nivel mundial se encuentran sobre el modelo relacional.

Las bases de datos no relacionales, incluyendo las jerárquicas, orientadas a grafos y orientadas a objetos existen desde finales de la década de 1960 (Leavitt, 2010). Con el surgimiento y desarrollo de la Web 3.0, las bases de datos orientadas a grafos que soportan el modelo de datos RDF han ganado popularidad (Zaki y cols., 2017; Brisaboa y cols., 2017). El lenguaje "Protocolo SPARQL y Lenguaje de Consulta RDF" ("SPARQL Protocol and RDF Query Language" - SPARQL, por sus siglas de acuerdo al término en idioma inglés) es la recomendación del "Consorcio de la World Wide Web" ("World Wide Web Consortium" - W3C, por sus siglas de acuerdo al término en idioma inglés) para la consulta sobre grafos RDF (W3C SPARQL Working Group, 2013).

Cada vez son más los servicios que se publican en la Web usando la "Transferencia de estado representacional" ("Representational state transfer" - REST, por sus siglas de acuerdo al término en idioma inglés) (Pautasso, 2014). El estilo de arquitectura REST enfatiza la escalabilidad de las interacciones entre los componentes de las aplicaciones informáticas y promueve la reutilización de los componentes reduciendo su acoplamiento (Pautasso, 2014). Resulta posible considerar aplicaciones externas que exponen sus datos a través de REST como fuentes de datos. En este caso, el análisis estructural de la fuente de datos se centraría en la estructura sintáctica que debe poseer la petición a formular para extraer los datos a integrar.

El Experto de dominio en este paso analiza cómo están físicamente distribuidos los datos que se pretenden integrar en las fuentes de datos que los almacenan. Posteriormente formula las consultas sintácticas que permitirán recuperar dichos datos utilizando los constructos definidos para este fin en las fuentes de datos a encuestar.

2.2.2. Análisis semántico de la información almacenada en cada fuente de datos

En el análisis semántico de la información almacenada en cada fuente de datos se determinan los conceptos contenidos en las mismas cuyos datos asociados se desean recuperar. En este paso interviene un Experto del dominio y un Ingeniero de conocimiento. El Ingeniero de conocimiento tiene la función de traducir los elementos presentes en una fuente de datos a conceptos comprensibles por un usuario externo.

En este paso un Usuario de conceptualización requiere ciertos conceptos sobre un dominio específico, esta petición se la realiza a un Experto del dominio. El Experto del dominio determina exactamente qué elementos del dominio son los que necesita el Usuario de conceptualización y se lo informa al Ingeniero de conocimiento. El Ingeniero de conocimiento modela conceptualmente los elementos recibidos en términos conocidos por el Usuario de conceptualización y retorna la conceptualización elaborada. Este proceso se ilustra en la figura 2.3.

Proceso de desarrollo de la ontología según la metodología METHONTOLOGY

El proceso de desarrollo de una ontología se refiere a cuáles actividades deben realizarse al construir una ontología (Gómez-Pérez y cols., 2007). La metodología METHONTOLOGY propone la realización de tres categorías de actividades que se ilustran en la figura 2.4.

Las actividades de gestión incluyen la planificación, el control y el aseguramiento de la calidad. La actividad de *planificación* identifica las tareas a desarrollar, su prioridad, el tiempo y los recursos para llevarlas a cabo. El resultado de esta actividad se ilustra en la figura 2.5. La actividad de *control* garantiza que las tareas planificadas se completen conforme a la manera en que se diseñaron. Finalmente, la actividad de *aseguramiento de la calidad* verifica que la calidad de cada artefacto generado (la ontología, la aplicación informática y la documentación) sea satisfactoria (Gómez-Pérez y cols., 2007). El cumplimiento de la actividad de *aseguramiento de la calidad* fue certificado por el tribunal que evaluó el ejercicio de culminación de estudios titulado "AUCTORITAS 2.0: Sistema de apoyo para el control de autoridades".

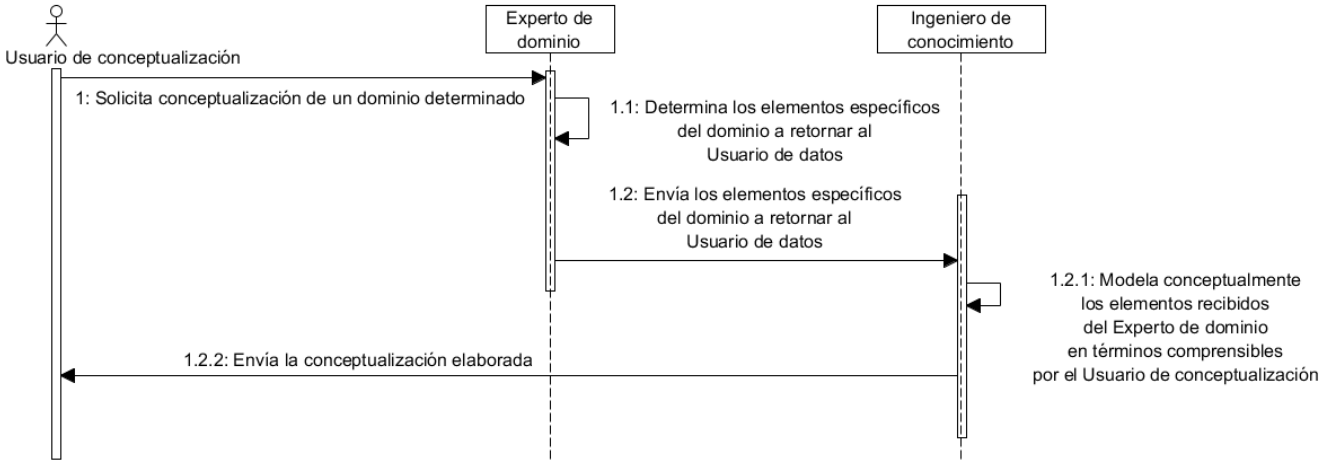

Figura 2.3: Secuencia realizada en el paso Análisis semántico de la información almacenada en cada fuente de datos

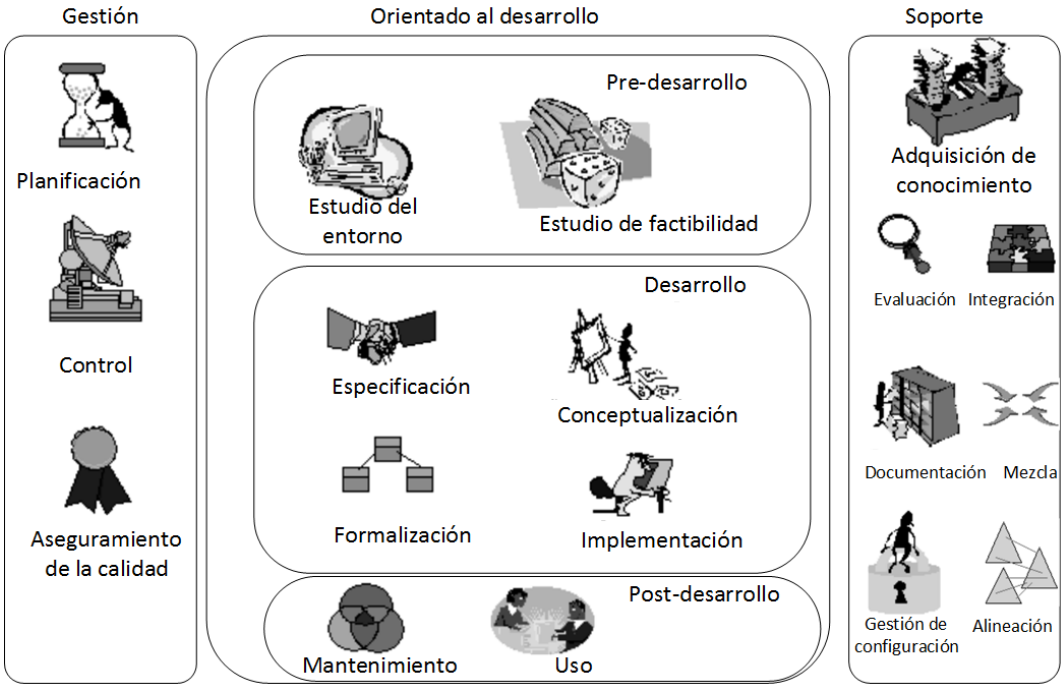

Figura 2.4: Proceso de desarrollo de una ontología. Tomado de (Gómez-Pérez y cols., 2007)

Task Name	Duration	Start	Finish	Predecessors	Resource Names
Identificación de los componentes básicos de una ontología para la OBDI	3 days	Mon 01/02/16	Wed 03/02/16		Leandro Tabares Martín
Identificación de los componentes requeridos para realizar OBDI en el contexto de aplicación	6 days	Thu 04/02/16	Thu 11/02/16	1	Leandro Tabares Martín
Identificar modificaciones a realizar en la aplicación informática donde se utilizará la ontología	7 days	Fri 12/02/16	Mon 22/02/16	2	Leandro Tabares Martín
Realizar el análisis de factibilidad de la ontología	4 days	Tue 23/02/16	Fri 26/02/16	3	Leandro Tabares Martín
Identificar los conceptos relevantes a incluir en la ontología	5 days	Mon 29/02/16	Fri 04/03/16	4	Leandro Tabares Martín
Formalizar la conceptualización elaborada	3 days	Mon 07/03/16	Wed 09/03/16	5	Leandro Tabares Martín
Construir la ontología utilizando el Lenguaje de Ontologías Web	7 days	Thu 10/03/16	Fri 18/03/16	6	Leandro Tabares Martín

Figura 2.5: Planificación de las tareas para el desarrollo de la ontología

Las actividades orientadas al desarrollo de la ontología se agrupan en actividades pre-desarrollo, actividades de desarrollo y actividades post-desarrollo. Durante las actividades pre-desarrollo se realiza un *estudio del entorno* con el fin de conocer las plataformas en las que se utilizará la ontología, las aplicaciones en las que la ontología será integrada, entre otros detalles. Como resultado de esta actividad se determinó que la ontología se utilizaría en una plataforma Java, al ser integrada en la aplicación informática AUCTORITAS versión 2.0. También, durante las actividades pre-desarrollo se lleva a cabo un *análisis de factibilidad* que debe responder a preguntas como: ¿Es posible construir la ontología? ¿Es factible construir la ontología? (Gómez-Pérez y cols., 2007).

Una vez en el desarrollo, la actividad *especificación* determina el por qué la ontología se está construyendo, para qué se usará y quiénes son los usuarios finales. La *conceptualización* estructura el dominio de conocimiento en forma de modelos significativos al nivel de conocimiento, el resultado de esta actividad se ilustra en la figura 2.6.

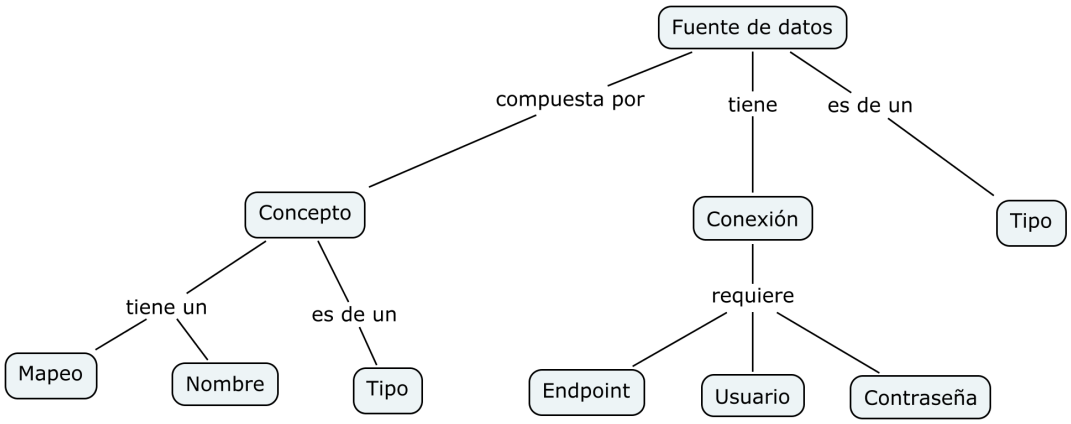

Figura 2.6: Conceptualización de los elementos del dominio de conocimiento

La actividad *formalización* transforma el modelo conceptual en un modelo formal o semi-computable, por medio de la utilización de los constructos definidos en la Lógica Descriptiva (Baader y cols., 2003). Para formalizar la conceptualización elaborada se decidió desarrollar la ontología que se ilustra en la figura 2.7. Esta ontología cumple con el requisito definido por Calvanese y cols. (2017), el cual especifica que los elementos básicos que debe poseer una ontología diseñada con el propósito de permitir la OBDI son los definidos en la ontología 2.1.

$$Concept \equiv Thing \sqcap \exists mappedTo.Literal \sqcap = 1 mappedTo \sqcap \exists type.Literal \sqcap = 1 type \sqcap \exists name.String \sqcap = 1 name$$

$$Connection \equiv Thing \sqcap \exists endpoint.Literal \sqcap = 1 endpoint \sqcap \forall user.String \sqcap \forall password.String$$

$$Datasource \equiv Thing \sqcap \exists composedBy.Concept \sqcap \exists has.Connection \sqcap = 1 has \sqcap type.Literal \sqcap = 1 type$$

Figura 2.7: T-Box de la ontología desarrollada

$$Concepto \equiv Thing \sqcap \exists mapeo.Literal \tag{2.1}$$

La actividad de *implementación* construye modelos computables en un lenguaje de ontologías (Gómez-Pérez y cols., 2007). Este modelo se construyó utilizando el Lenguaje de Ontologías Web por medio de la herramienta Protégé y se ilustra en la figura 2.8.

Durante el post-desarrollo, la actividad de *mantenimiento* actualiza y corrige la ontología de ser necesario. También en esta fase la ontología es reutilizada por otras ontologías o aplicaciones (Gómez-Pérez y cols., 2007).

Finalmente, las actividades de soporte incluyen una serie de actividades que se realizan al mismo tiempo que las orientadas al desarrollo. Ellas incluyen la adquisición de conocimiento, evaluación, integración, mezcla, alineación,

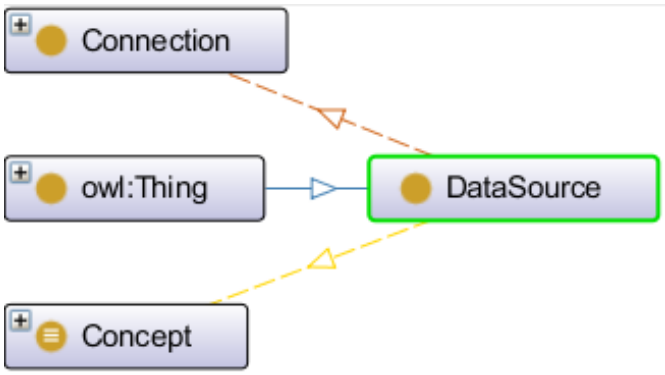

Figura 2.8: Clases de la ontología construida con la herramienta Protégé

documentación y gestión de configuración. El objetivo de la actividad de *adquisición de conocimiento* es obtener el conocimiento de expertos del dominio a través de algún tipo de procedimiento (semi)automático, lo que se conoce como aprendizaje ontológico. La actividad de *evaluación* realiza un juicio técnico de las ontologías, sus entornos de aplicaciones informáticas asociados, así como de la documentación. Este juicio se realiza con respecto a un marco de referencia durante cada etapa y entre etapas durante el ciclo de vida de la ontología. La actividad de *integración* se requiere cuando se construye la ontología a partir de la reutilización de otras ontologías (Gómez-Pérez y cols., 2007).

Otra actividad de soporte es la *mezcla*, que consiste en la obtención de una ontología a partir de diferentes ontologías del mismo dominio. La actividad de *alineación* establece diferentes tipos de mapeos entre las ontologías involucradas. Por otra parte, la actividad de *documentación* detalla, clara y exhaustivamente cada una de las etapas completadas y los productos generados. La *gestión de configuración* almacena todas las versiones de la documentación y del código de la ontología para controlar los cambios.

2.2.3. Instanciación de la ontología para la integración de datos

Jannach y cols. (2009) han demostrado que las ontologías de dominio pueden apoyar la extracción de conocimiento, lo que significa que el problema de extracción de información puede ser generalizado al problema de encontrar e insertar información que concuerde con una ontología en la base de conocimiento de un sistema, este proceso se conoce como "instanciación de la ontología" o "poblado de la ontología". Entonces la ontología instanciada puede servir como base para el desarrollo de servicios de conocimiento en la Web Semántica (Alani y cols., 2003).

La instanciación de una ontología es un proceso costoso y que consume tiempo (Makki y cols., 2008), por esta razón existen enfoques automáticos (Alani y cols., 2003; Jannach y cols., 2009) y semi-automáticos (Makki y cols., 2008). La instanciación de la ontología para OBDI consiste en la representación de los conceptos identificados en

Herramienta	Versión	Propietario / Desarrollador	Características / Limitaciones	Lenguaje primario	Libre Libre
Adaptiva	-	Universidad Sheffield	Adquisición de conocimiento	Java	Sí
Semantic-works 2012	2012	Altova	Editor OWL + RDFS	Java	No
Conzilla	2.2	Grupo de investigaciones de gestión del conocimiento	Navegador de conceptos	Java	Sí
HOZO	5.6	Universidad de Osaka	Amigable al usuario	Java	Sí
Protégé	5.5.0	Universidad de Stanford	Herencia múltiple	Java	Sí

Tabla 2.2: Editores de ontologías

el paso anterior acorde a la terminología definida por la ontología seleccionada. Debido al alto nivel de detalle que implica instanciar una ontología para OBDI se escogió realizar este proceso manualmente.

Para la instanciación manual de una ontología es posible apoyarse en un editor de ontologías. La tabla 2.2 sintetiza características de varios editores de ontologías.

Un ejemplo de instanciación se ilustra en la figura 2.9, en la que se utilizó la variable syntacticQuery con el objetivo de facilitar su comprensión. Esta variable se refiere a una consulta sintáctica formulada en un lenguaje de consulta (por ejemplo: SQL, SPARQL, una petición REST, entre otros).

$Datasource(PostgreSQL),\ has(PostgreSQL, postgresConnection),\ type(PostgreSQL, local)$
$composedBy(PostgreSQL, autorPersonalPostgres).$
$Concept(autorPersonalPostgres),\ mappedTo(autorPersonalPostgres, syntacticQuery),$
$type(autorPersonalPostgres, AUTORPERSONAL),\ name(autorPersonalPostgres, Autor\ Personal).$
$Connection(postgresConnection),\ user(postgresConnection, postgres),$
$password(postgresConnection, admin),$
$endpoint(postgresConnection, jdbc{:}postgresql{:}//localhost{:}5432/Autoridades).$
$Datasource(VIVO),\ has(VIVO, vivoConnection),\ composedBy(VIVO, institution),$
$composedBy(VIVO, autorPersonalVivo),\ type(PostgreSQL, http).$
$Concept(institution),\ mappedTo(institution, syntacticQuery),\ name(institution, Institution)$
$type(institution, AUTORCORPORATIVO).$
$Concept(autorPersonalVivo),\ mappedTo(autorPersonalVivo, syntacticQuery),$
$type(autorPersonalVivo, AUTORPERSONAL), name(autorPersonalVivo, Autor\ Personal).$
$Connection(vivoConnection),$
$endpoint(vivoConnection,$
$http{:}//localhost{:}8080/vivo/api/sparqlQuery?email{=}vivoroot@mydomain.edu\&password{=}zas123..).$
$Datasource(ORCID),\ composedBy(ORCID, autorPersonalOrcid),\ has(ORCID, orcidConnection),$
$type(ORCID, orcid).$
$Concept(autorPersonalOrcid),\ mappedTo(autorPersonalOrcid, syntacticQuery),$
$type(autorPersonalOrcid, AUTORPERSONAL),\ name(autorPersonalOrcid, Autor\ Personal.)$
$Connection(orcidConnection),$
$endpoint(orcidConnection, https{:}//pub.orcid.org/v1.2/search/orcid{-}bio/?q{=}).$
$Datasource(AGROVOC),\ composedBy(AGROVOC, agrovocConcept),$
$has(AGROVOC, virtuosoConnection),\ type(AGROVOC, http).$
$Concept(agrovocConcept),\ mappedTo(agrovocConcept, syntacticQuery),$
$type(agrovocConcept, CONTROLEDTERMS),\ name(agrovocConcept, Termino\ Controlado)$
$Connection(virtuosoConnection), endpoint(virtuosoConnection, http{:}//localhost{:}8890/sparql).$
$Datasource(ACM),\ composedBy(ACM, acmConcept),\ has(ACM, virtuosoConnection),$
$type(ACM, http).$
$Concept(acmConcept),\ mappedTo(acmConcept, syntacticQuery),$
$type(acmConcept, CONTROLEDTERMS),\ name(acmConcept, Termino\ Controlado)$

Figura 2.9: A-Box de la ontología desarrollada

2.2.4. Publicación de la instancia de la ontología creada

Uno de los principales retos que se enfrentan en la Web Semántica es que las ontologías típicamente se publican como ficheros sin soporte para utilizarlos en aplicaciones (Viljanen y cols., 2008). Los almacenes de tripletas RDF constituyen un tipo de bases de datos basadas en grafos, lo que les permite almacenar datos en forma de redes de objetos vinculados entre sí. Las bases de datos de este tipo son capaces de gestionar consultas semánticas y utilizar inferencias para descubrir nueva información implícita en los datos almacenados y sus relaciones.

La publicación de la instancia de la ontología creada se refiere a su carga en un almacén de tripletas RDF. Este paso permite compartir en línea la instancia, haciéndola reutilizable por otras aplicaciones que requieran acceder a las fuentes de datos descritas. Los elementos descritos en una ontología publicada en un almacén de tripletas RDF son accesibles mediante consultas SPARQL, beneficiándose de las características de este tipo de bases de datos y de las potencialidades del lenguaje de consultas.

2.2.5. Consumo por una aplicación informática de la instancia de la ontología creada

El consumo por una aplicación informática de la instancia de la ontología creada permite, a partir de la descripción de los términos de la ontología, obtener su instanciación y utilizarla para recuperar los datos requeridos. En este paso se utiliza la T-Box de la ontología como base para la interpretación semántica de la estructura de la fuente de datos a encuestar.

La A-Box de la ontología brinda a la aplicación informática los datos necesarios para gestionar la conexión a la fuente de datos. A su vez, contiene los conceptos relevantes a recuperar y las consultas sintácticas que posibilitan este proceso.

2.3. Conclusiones del capítulo

El paradigma de las ciencias del diseño permitió definir el artefacto a generar como aporte teórico principal de la investigación. El proceso utilizado para la investigación en ciencias del diseño condujo las actividades a realizar como parte del presente trabajo.

El método creado permite la integración de datos almacenados en fuentes estructuralmente heterogéneas, brindando una guía a los especialistas encargados de este proceso. La utilización de la metodología para el desarrollo de ontologías seleccionada, orientó satisfactoriamente el proceso de desarrollo de la ontología que se propone como parte del método.

Capítulo 3

Validación de la propuesta

3.1. Introducción

En este capítulo de presenta la validación del método para la integración de datos basado en ontologías OntoIntegra. Para la validación se utilizó un caso de estudio sobre dos versiones de la aplicación informática AUCTORITAS. En él se presenta el diseño del caso de estudio y se realiza un análisis de los resultados obtenidos concluyendo con la confirmación de la hipótesis planteada en la investigación.

3.2. Selección de la estrategia de validación

En el paradigma de las ciencias del diseño, Hevner y cols. (2004) plantean en la "Guía 3: Evaluación del diseño" que la utilidad, calidad y eficacia de un artefacto de diseño debe ser rigurosamente demostrada a través de métodos de evaluación bien ejecutados. La evaluación de un artefacto tecnológico requiere la definición de métricas apropiadas y posiblemente la compilación y análisis de datos apropiados (Hevner y cols., 2004). La tabla 3.1 sintetiza los métodos de evaluación de diseño propuestos por Hevner y cols. (2004).

Dependiendo del propósito de la evaluación y de las condiciones de la investigación empírica, existen tres grandes tipos de estrategias que pueden utilizarse: encuestas, casos de estudio y experimentos (Wohlin y cols., 2012).

Una encuesta es un sistema para recolectar información sobre o acerca de personas que describen, comparan o explican su conocimiento, actitudes y comportamiento (Wohlin y cols., 2012). Una encuesta no permite el control sobre la ejecución de la medición, por lo que no es posible manipular variables como en otros métodos de investigación (Wohlin y cols., 2012).

Los casos de estudio en la ingeniería de software son investigaciones empíricas que se basan en múltiples fuentes de evidencia para investigar una instancia (o un pequeño número de ellas) de un fenómeno de ingeniería de software contemporáneo dentro de su contexto real, especialmente cuando el límite entre el fenómeno y el contexto no puede ser claramente definido (Runeson y cols., 2012; Wohlin y cols., 2012). Constituyen una técnica donde factores clave que pueden incidir en la salida se identifican y se documenta la actividad (Stake, 1995).

Tipo de método	Método
Observacional	Caso de estudio: artefacto de estudio en profundidad del entorno del negocio. Estudio de campo: supervisa el uso de un artefacto en varios proyectos.
Analítico	Análisis estático: examina las cualidades estáticas de la estructura del artefacto. Análisis de Arquitectura: estudia cómo corresponde el artefacto dentro de la arquitectura del SI. Optimización: demuestra propiedades óptimas inherentes al artefacto o provee límites óptimos en el comportamiento del artefacto. Análisis dinámico: estudia las cualidades dinámicas del artefacto.
Experimental	Experimento controlado: estudia características del artefacto en un ambiente controlado. Simulación: ejecuta el artefacto con datos artificiales.
Prueba	Prueba funcional (de caja negra): ejecuta las interfaces del artefacto para descubrir fallos e identificar defectos. Prueba estructural (de caja blanca): realiza pruebas de cubrimiento de alguna métrica en la implementación del artefacto.
Descriptivo	Argumento informado: usa información proveniente de la base de conocimiento para construir un argumento convincente sobre la utilidad del artefacto. Escenarios: construye escenarios detallados alrededor del artefacto para demostrar su utilidad.

Tabla 3.1: Métodos de evaluación del diseño. Tomado de (Hevner y cols., 2004).

Los experimentos (o experimentos controlados) en la ingeniería de software son un tipo de investigación empírica que manipula un factor o variable de la configuración estudiada. Basado en la aleatoriedad se aplican diferentes tratamientos a diferentes sujetos, mientras se mantienen otras variables constantes y se miden los efectos en las variables de salida (Wohlin y cols., 2012). Constituyen una investigación formal, rigurosa y controlada en la que los factores claves son identificados y manipulados, mientras que otros factores en el contexto se mantienen sin cambio.

Los quasi-experimentos constituyen un tipo de investigación empírica similar a un experimento, donde la asignación de los tratamientos a sujetos no puede ser aleatorizada, sino que emerge de las características propias de los sujetos u objetos (Wohlin y cols., 2012).

La diferencia entre casos de estudio y experimentos está determinada por el nivel de control del contexto (Petersen y Wohlin, 2009). En un experimento diferentes situaciones son forzadas deliberadamente y el objetivo comúnmente es distinguir entre las dos situaciones. En un caso de estudio el contexto es controlado por el proyecto real analizado (Wohlin y cols., 2012). En el contexto del proyecto ELINF, las fuentes de datos a utilizar para el control de autoridades se incrementarán en la medida de las necesidades. Una vez analizado el contexto en que se utilizará la aplicación informática AUCTORITAS, el autor de la presente investigación concluye que la realización de un caso de estudio constituye una estrategia viable para la validación de la misma, ya que se realizará la integración de las fuentes de datos para el control de autoridades relevantes al proyecto ELINF.

3.3. Preparación del caso de estudio

Según Kitchenham y cols. (1995) para evitar el sesgo y asegurar la validez interna, es necesario crear una base sólida para evaluar los resultados de un caso de estudio. Kitchenham y cols. (1995) proponen tres alternativas para preparar un estudio:

- Una comparación de los resultados aplicando el método contra una línea base.
- Un proyecto hermano puede ser seleccionado como línea base. El proyecto bajo estudio emplea el nuevo método mientras que el proyecto hermano usa los métodos anteriores. Ambos proyectos deben ser comparables.
- Si el método se aplica a componentes del producto individuales, debe ser aplicado aleatoriamente a algunos componentes y a otros no.

Para el caso de estudio en cuestión se considera la segunda alternativa de las propuestas por Kitchenham y cols. (1995), por lo que se utilizarán las versiones 1.0 y 2.0 de la aplicación informática AUCTORITAS.

Según Wohlin y cols. (2012), la realización de un caso de estudio involucra cinco grandes pasos por los que transitar:

1. Diseño del caso de estudio: se definen los objetivos y se planifica el caso de estudio.
2. Preparación para la recolección de los datos: se definen los procedimientos y protocolos para la recolección de los datos.

3. Recolección de los datos: ejecución de la recolección de los datos en el caso estudiado.

4. Análisis de los datos recolectados.

5. Reporte del caso de estudio.

3.3.1. Diseño del caso de estudio

El **objetivo** del caso de estudio es **medir qué por ciento de las fuentes de datos, necesarias para el control de autoridades en el proyecto ELINF, es posible integrar mediante el empleo del método propuesto en una aplicación informática**.

El presente caso de estudio se centra en la escalabilidad en cuanto a fuentes de datos de la aplicación informática AUCTORITAS. Se asume como escalabilidad el concepto elaborado por Duboc y cols. (2006) que la define como la cualidad de las aplicaciones informáticas caracterizada por el impacto causal que poseen aspectos del entorno del sistema según estos son variados por encima de los rangos operacionales.

Las preguntas de investigación que conducirán el presente caso de estudio son:

1. ¿Qué fuentes de datos de las necesarias para el control de autoridades en el proyecto ELINF es posible integrar con el método propuesto?

2. ¿Qué nivel de flexibilidad aporta la aplicación del método propuesto en el acceso a datos de la aplicación informática AUCTORITAS?

Se definen los siguientes umbrales para la variable flexibilidad:

- Flexibilidad baja: es necesario modificar el código fuente de la aplicación para adicionar nuevas fuentes de datos y estas deben ser estructuralmente homogéneas.

- Flexibilidad media: es necesario modificar el código fuente de la aplicación para adicionar nuevas fuentes de datos pero estas pueden ser estructuralmente heterogéneas.

- Flexibilidad alta: no es necesario modificar el código fuente de la aplicación para adicionar nuevas fuentes de datos y estas pueden ser estructuralmente heterogéneas.

El Comité de Expertos del proyecto ELINF identificó como fuentes de datos necesarias para el control de autoridades las siguientes:

- Fichero de autoridad local: base de datos que contiene las entradas de autoridad de autores cubanos que no están registrados en fuentes internacionales.

- VIVO: aplicación informática que gestiona perfiles de investigadores. Esta aplicación se pretende implementar en la red de universidades miembros del proyecto ELINF.
- ORCID: fuente de datos internacional para la identificación de autores.
- Tesauro de la ACM: tesauro especializado en Ciencias de la Computación para el control de autoridades a nivel de epígrafes de materia.
- Tesauro AGROVOC de la FAO: tesauro especializado en Ciencias Agrícolas para el control de autoridades a nivel de epígrafes de materia.

3.3.2. Recolección de los datos

De acuerdo con Lethbridge y cols. (2005) las técnicas para la recolección de datos se pueden dividir en tres niveles:

- Primer nivel: Métodos en los que el investigador está en contacto directo con los sujetos y recolecta los datos en tiempo real.
- Segundo nivel: Métodos indirectos en los que el investigador recolecta datos en bruto sin interactuar con los sujetos durante la recolección.
- Tercer nivel: Análisis independiente de artefactos de trabajo donde se utilizan datos disponibles y en algunos casos ya compilados.

La técnica para recolectar datos en la presente investigación cae en el tercer nivel, basándose en las investigaciones de Calzadilla-Reyes y Ruano-Alvarez (2015) y González-Barroso y Pérez-González (2016).

El resultado de la recolección de datos se muestra en la tabla 3.2.

3.3.3. Análisis de los datos recolectados

El principal objetivo del análisis cualitativo de los datos es llegar a conclusiones a partir de los datos, manteniendo una clara cadena de evidencia (Wohlin y cols., 2012). Existen dos partes diferentes de análisis sobre datos cualitativos: las técnicas generadoras de hipótesis y las técnicas para confirmación de hipótesis (Seaman, 1999). Las técnicas generadoras de hipótesis pretenden arribar a hipótesis a partir de los datos, mientras que las técnicas para confirmación de hipótesis se utilizan para demostrar que una hipótesis es verdadera. El análisis de los datos recolectados en el presente caso de estudio pretenderá demostrar la validez de la hipótesis de la investigación.

Aplicación informática	Fuente de datos	Estructura	Modificación código fuente
AUCTORITAS 1.0	Fichero de autoridad local	Modelo relacional	No
AUCTORITAS 1.0	VIVO	Servicio REST	Sí
AUCTORITAS 1.0	ORCID	Servicio REST	Sí
AUCTORITAS 1.0	Tesauro ACM	Modelo RDF	Sí
AUCTORITAS 1.0	Tesauro AGROVOC	Modelo RDF	Sí
AUCTORITAS 2.0	Fichero de autoridad local	Modelo relacional	No
AUCTORITAS 2.0	VIVO	Servicio REST	No
AUCTORITAS 2.0	ORCID	Servicio REST	Sí
AUCTORITAS 2.0	Tesauro ACM	Modelo RDF	No
AUCTORITAS 2.0	Tesauro AGROVOC	Modelo RDF	No

Tabla 3.2: Datos recolectados en el caso de estudio.

El análisis de los datos puede realizarse con diferentes niveles de formalismo, Wohlin y cols. (2012) mencionan los siguientes enfoques:

- Enfoques de inmersión: Son los enfoques menos estructurados, más dependientes de la intuición y las habilidades interpretativas del investigador. Pueden ser difíciles de combinar con los requerimientos para mantener y comunicar la cadena de evidencia.
- Enfoques de edición: Incluyen pocos códigos pre-elaborados.
- Enfoques basados en plantillas: Son más formales e incluyen preguntas de investigación creadas a priori.
- Enfoques quasi-estadísticos: Son altamente formales e incluyen, por ejemplo, cálculos o frecuencias de palabras o frases.

El análisis de los datos en el presente caso de estudio utilizará un enfoque basado en plantillas.

La versión 1.0 de la aplicación informática AUCTORITAS fue desarrollada con una clase de acceso a datos, en la cual se gestionaba el acceso a una base de datos relacional conteniendo los registros de autoridades locales. La lógica de acceso a datos estaba altamente acoplada al consumo de datos a partir de un modelo relacional, por lo que incorporar nuevas fuentes con un modelo de datos diferente provocaba cambios en la lógica de acceso a datos de la aplicación.

El acceso a datos de la versión 2.0 de AUCTORITAS de desarrolló siguiendo el método propuesto en el presente trabajo. La lógica de acceso a datos se hizo dependiente de la parte conceptual de la ontología creada, mientras

que las fuentes de datos son configuradas en la instanciación de dicha ontología, constituyendo su A-Box. Esto posibilitó que la integración de nuevas fuentes de datos no produjese cambios en el código fuente de la aplicación excepto en un caso.

En el caso de la fuente de datos ORCID fue necesario modificar el código fuente de la aplicación, debido a que esta fuente utiliza un mecanismo de autenticación OAuth 2.0 que no había sido contemplado en el desarrollo de AUCTORITAS. Teniendo esto en cuenta, se realizó la descripción de la fuente de datos en la instancia de ontología, se incorporó el mecanismo de autenticación a la aplicación y las pruebas resultaron satisfactorias.

La realización del caso de estudio permitió identificar que la versión 1.0 de AUCTORITAS posee una flexibilidad media en su acceso a datos, mientras que la aplicación del método propuesto en el desarrollo de la versión 2.0 aportó un nivel de flexibilidad alto. De igual manera, fue posible verificar que el método propuesto permite integrar el cien por ciento de las fuentes de datos necesarias para el control de autoridades en el proyecto ELINF. Los resultados arrojados por el caso de estudio llevado a cabo comprueban la validez de la hipótesis propuesta en la presente investigación.

3.4. Conclusiones del capítulo

La realización del caso de estudio da lugar a las siguientes conclusiones respecto a la validación del método propuesto:

- La herramienta informática desarrollada como instanciación del método propuesto facilitó la obtención de información requerida para el proceso de control de autoridades en el proyecto ELINF.
- La utilización de Integración de Datos Basada en Ontologías, disminuyó el acoplamiento de la herramienta informática desarrollada como instanciación del método propuesto, con la configuración de las fuentes de datos requeridas para el proceso de control de autoridades en el proyecto ELINF.
- La instanciación del método propuesto en el acceso a datos de la aplicación informática AUCTORITAS permitió integrar la totalidad de las fuentes de datos requeridas por el proyecto ELINF para el proceso de control de autoridades.

Conclusiones generales

La realización de la presente investigación ratificó la necesidad de un método que conduzca el proceso de integración de datos almacenados en fuentes heterogéneas de datos y que aporte semántica al proceso. A su vez, se confirmó la Integración de Datos Basada en Ontologías como una de las vías más promisorias para la realización del proceso de integración de datos. Por otra parte, la utilización del paradigma de las ciencias del diseño resultó efectiva en el desarrollo del método propuesto. El método OntoIntegra permitió la integración de datos almacenados en fuentes estructuralmente heterogéneas, mientras que el caso de estudio utilizado para la validación del método propuesto certificó su efectividad, por medio de la instanciación de la propuesta en una aplicación informática.

Recomendaciones

Para futuras investigaciones y como continuidad de la actual se recomienda:

- Enriquecer la semántica de la ontología desarrollada, para contribuir al descubrimiento de información implícita en las relaciones entre los datos.
- Generalizar el método propuesto a otras aplicaciones que requieran integrar datos almacenados en fuentes heterogéneas.

Referencias bibliográficas

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., y Shadbolt, N. R. (2003, Jan). Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, *18*(1), 14-21. doi: 10 .1109/MIS.2003.1179189

Alfred, R., Chung, C. J., On, C. K., Ibrahim, A. A. A., Sainin, M. S., y Pandiyan, P. M. (2018). Data fusion based on self-organizing map approach to learning medical relational data. En R. Alfred, H. Iida, A. A. Ag. Ibrahim, y Y. Lim (Eds.), *Computational science and technology* (pp. 230–240). Singapore: Springer Singapore.

Alooma Inc. (2017). *ETL software tools.* Descargado 2018-01-15, de https://www.etltools.net/

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., y Patel-Schneider, P. F. (Eds.). (2003). *The Description Logic Handbook*. New York, NY, USA: Cambridge University Press.

Berners-Lee, T. (2006). *Linked Data.* Descargado 12/28/2017, de http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J., y Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34–43. doi: 10.1038/scientificamerican0501-34

Berners-Lee, T. J., y Cailliau, R. (1990). Worldwideweb: Proposal for a hypertext project.

Bregzis, R. (1982). The syndetic structure of the catalog. *Authority control: the key to tomorrow's catalog. Proceedings of the 1979 Library and Information Technology Association Institute, Mary W. Ghikas ed. Phoenix: AZ*.

Brisaboa, N. R., Cerdeira-Pena, A., de Bernardo, G., y Navarro, G. (2017). Compressed representation of dynamic binary relations with applications. *Information Systems*, *69*, 106 - 123. Descargado de http://www.sciencedirect.com/science/article/pii/S030643791630535X doi: https://doi.org/10.1016/j.is.2017.05.003

Buckles, B. P., y Petry, F. E. (1993). A fuzzy representation of data for relational databases. En D. Dubois, H. Prade, y R. R. Yager (Eds.), *Readings in fuzzy sets for intelligent systems* (p. 660 - 666). Morgan Kaufmann. Descargado de https://www.sciencedirect.com/science/article/pii/B9781483214504500717 doi: https://doi.org/10.1016/B978-1-4832-1450-4.50071-7

Calvanese, D., Cogrel, B., Komla-Ebri, S., Lanti, D., Rezk, M., y Xiao, G. (2015). How to Stay Ontop of Your Data: Databases, Ontologies and More. En *Revised selected papers of the eswc 2015 satellite events on the semantic web: Eswc 2015 satellite events - volume 9341* (pp. 20–25). New York, NY, USA: Springer-Verlag New York, Inc. Descargado de http://dx.doi.org/10.1007/978-3-319-25639-9{_}4 doi: 10.1007/978-3-319-25639-9_4

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., y Rosati, R. (2017). Ontology-Based Data Access and Integration. En L. Liu y M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 1–7). New York, NY: Springer New York. Descargado de https://doi.org/10.1007/978-1-4899-7993-3{_}80667-1 doi: 10.1007/978-1-4899-7993-3_80667-1

Calvanese, D., Liuzzo, P., Mosca, A., Remesal, J., Rezk, M., y Rull, G. (2016). Ontology-based data integration in EPNet: Production and distribution of food during the Roman Empire. *Engineering Applications of Artificial Intelligence*, *51*, 212–229. Descargado de http://www.sciencedirect.com/science/article/pii/S0952197616000099 doi: https://doi.org/10.1016/j.engappai.2016.01.005

Calzadilla-Reyes, D., y Ruano-Alvarez, W. A. (2015). *AUCTORITAS Sistema de apoyo para el control de autoridades* (Trabajo de diploma). Universidad de las Ciencias Informáticas.

Carrasco, R. C., Serrano, A., y Castillo-Buergo, R. (2016). A parser for authority control of author names in bibliographic records. *Information Processing and Management*, *52*(5), 753–764. doi: 10.1016/j.ipm.2016.02.002

Ciudad-Ricardo, F. A., Goovaerts, M., Meneses-Placeres, G., Batista-Matamoros, C. R., Leiva-Mederos, A., Machado-Rivero, M. O., ... Díaz-Pérez, M. (2017). *Self-assessment NETWORK Partnership Project level ICT Supporting the educational processes and the knowledge management in higher education (ELINF)* (Inf. Téc.). La Habana.

Corby, O., Gaignard, A., Zucker, C. F., y Montagnat, J. (2012). KGRAM Versatile Inference and Query Engine for the Web of Linked Data. En *Proceedings of the the 2012 ieee/wic/acm international joint conferences on web intelligence and intelligent agent technology - volume 01* (pp. 121–128). Washington, DC, USA: IEEE Computer Society. Descargado de http://dl.acm.org/citation.cfm?id=2457524.2457672

Cutter, C. A. (1889). *Rules for a printed dictionary catalogue*. US Government Printing Office.

Daraio, C., Lenzerini, M., Leporelli, C., Moed, H. F., Naggar, P., Bonaccorsi, A., y Bartolucci, A. (2016, feb). Data integration for research and innovation policy: an Ontology-Based Data Management approach. *Scientometrics*,

106(2), 857–871. Descargado de https://doi.org/10.1007/s11192-015-1814-0 doi: 10.1007/s11192-015-1814-0

Denning, P. J. (1997). A New Social Contract for Research. *Communications of the ACM*, *40*(2), 132–134. Descargado de http://doi.acm.org/10.1145/253671.253755 doi: 10.1145/253671.253755

Doan, A., Halevy, A., y Ives, Z. (2012). *Principles of data integration* (1st ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Duboc, L., Rosenblum, D. S., y Wicks, T. (2006). A framework for modelling and analysis of software systems scalability. En *Proceedings of the 28th international conference on software engineering* (pp. 949–952).

Eisenberg, A., y Melton, J. (1999). SQL: 1999, Formerly Known As SQL3. *SIGMOD Rec.*, *28*(1), 131–138. Descargado de http://doi.acm.org/10.1145/309844.310075 doi: 10.1145/309844.310075

El Kadiri, S., Grabot, B., Thoben, K.-d., Hribernik, K., Emmanouilidis, C., Von Cieminski, G., y Kiritsis, D. (2015). Current trends on ICT technologies for enterprise information. *Computers in Industry*. doi: 10.1016/j.compind.2015.06.008

Elsevier. (2016). *Scopus*. Descargado de https://www.elsevier.com/solutions/scopus

Franke, M., Klein, K., Hribernik, K., Lappe, D., Veigt, M., y Thoben, K.-d. (2014). Semantic Web Service Wrappers as a foundation for interoperability in closed-loop Product Lifecycle Management. *Procedia CIRP*, *22*, 225–230. doi: 10.1016/j.procir.2014.07.020

García Noguera, M. (2009). *Modelado Y Análisis De Sistemas Cscw Siguiendo Un Enfoque De Ingeniería Dirigida Por Ontologías* (Tesis doctoral, Universidad de Granada). Descargado de http://0-hera.ugr.es.adrastea.ugr.es/tesisugr/18014094.pdf

García Rodríguez, S. (2016). *Propuesta de normalización para el control de autoridades en la red TIC Cuba* (Trabajo de Diploma, Universidad Central "Marta Abreu"de Las Villas). Descargado de http://dspace.uclv.edu.cu/handle/123456789/6618

Gómez-Pérez, A., Fernández-López, M., y Corcho, O. (2004). *Ontological Engineering* (1.ª ed.; X. Wu y L. Jain, Eds.). Londres: Springer-Verlag London.

Gómez-Pérez, A., Fernández-López, M., y Corcho, O. (2007). *Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. (Advanced Information and Knowledge Processing)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

González-Barroso, F., y Pérez-González, D. (2016). *AUCTORITAS 2.0: Sistema de apoyo para el Control de Autoridades* (Trabajo de diploma). Universidad de las Ciencias Informáticas.

Görlitz, O., y Staab, S. (2010). SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions. En *Proceedings of the second international conference on consuming linked data - volume 782* (pp. 13–24). Aachen, Germany, Germany: CEUR-WS.org. Descargado de http://dl.acm.org/citation.cfm?id=2887352.2887354

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. Descargado de http://linkinghub.elsevier.com/retrieve/pii/S1042814383710083 doi: 10.1006/knac.1993.1008

Gutierrez, C., Hurtado, C. A., Mendelzon, A. O., y Pérez, J. (2011). Foundations of Semantic Web databases. *Journal of Computer and System Sciences*, *77*(3), 520–541. doi: 10.1016/j.jcss.2010.04.009

H. Agus-Santoso, S. C.-H., y Abdul-Mehdi, Z. (2011). Ontology extraction from relational database: Concept hierarchy as background knowledge. *Knowledge-based Systems*, *24*(3), 457–464.

Han, H., Xu, W., Zha, H., y Giles, C. L. (2005). A hierarchical naive Bayes mixture model for name disambiguation in author citations. En *Proceedings of the 2005 acm symposium on applied computing - sac '05* (pp. 1065–1069). New Mexico: ACM. doi: 10.1145/1066677.1066920

Harper, C. A., y Tillett, B. B. (2007). Library of Congress Controlled Vocabularies and Their Application to the Semantic Web. *Cataloging & Classification Quarterly*, *43*(3/4), 47–68. doi: http://dx.doi.org/10.1300/J104v43n03_03

Heath, T., y Bizer, C. (2011). *Linked Data. Evolving the Web into a Global Data Space* (1.ª ed.; J. Hendler, Ed.). Morgan & Claypool. doi: 10.2200/S00334ED1V01Y201102WBE001

Hevner, A. R., March, S. T., Park, J., y Ram, S. (2004). Design Science in Information Systems Research. *MIS Q.*, *28*(1), 75–105. Descargado de http://dl.acm.org/citation.cfm?id=2017212.2017217

Hidalgo-Delgado, Y. (2015). *Marco de trabajo basado en los datos enlazados para la interoperabilidad semántica en el protocolo OAI-PMH* (Tesis de maestría). Universidad de las Ciencias Informáticas.

Hoang, H. H., Cung, T. N.-p., Truong, D. K., Hwang, D., y Jung, J. J. (2014). Semantic Information Integration with Linked Data Mashups Approaches. *International Journal of Distributed Sensor Networks*. doi: 10.1155/2014/813875

Hristidis, V., y Papakonstantinou, Y. (2002). Chapter 58 - discover: Keyword search in relational databases. En P. A. Bernstein y cols. (Eds.), *{VLDB} '02: Proceedings of the 28th international conference on very large databases* (p. 670 - 681). San Francisco: Morgan Kaufmann. Descargado de https://www.sciencedirect.com/science/article/pii/B9781558608696500652 doi: https://doi.org/10.1016/B978-155860869-6/50065-2

Ian Horrocks, P. F. P.-S., y van Harmelen, F. (2003). From SHIQ and RDF to OWL: The Making of a Web Ontology Language. *Web Semantics*, *1*(1), 7–26.

Ilyas, I. F., Aref, W. G., y Elmagarmid, A. K. (2004). Supporting top-k join queries in relational databases. *VLDB Journal*, *13*(3), 207–221. doi: 10.1007/s00778-004-0128-2

International Federation of Library Associations and Institutions. (2009). *Functional Requirements for Authority Data: A Conceptual Model* (Vol. 34; G. Patton, Ed.). La Haya, Holanda: International Federation of Library Associations and Institutions.

Iqbal, R., Azrifah, M., Murad, A., Mustapha, A., y Sharef, N. M. (2013). An Analysis of Ontology Engineering Methodologies : A Literature Review. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(16), 2993–3000.

ISNI International Standard Name Identifier. (2017). *ISNI*. Descargado 2017-12-15, de http://www.isni.org/

Janev, V., y Vranes, S. (2011). Applicability assessment of Semantic Web technologies. *Information Processing and Management*, *47*(4), 507–517. doi: 10.1016/j.ipm.2010.11.002

Jannach, D., Shchekotykhin, K., y Friedrich, G. (2009). Automated ontology instantiation from tabular web sources—The AllRight system. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7*(3), 136–153. Descargado de http://www.sciencedirect.com/science/article/pii/S1570826809000055 doi: https://doi.org/10.1016/j.websem.2009.04.002

Kharlamov, E., Brandt, S., Jimenez-Ruiz, E., Kotidis, Y., Lamparter, S., Mailis, T., ... Moeller, R. (2016). Ontology-Based Integration of Streaming and Static Relational Data with Optique. En *Proceedings of the 2016 international conference on management of data* (pp. 2109–2112). New York, NY, USA: ACM. Descargado de http://doi.acm.org/10.1145/2882903.2899385 doi: 10.1145/2882903.2899385

Kharlamov, E., Hovland, D., Skjæveland, M. G., Bilidas, D., Jiménez-Ruiz, E., Xiao, G., ... Waaler, A. (2017). Ontology Based Data Access in Statoil. *Web Semantics: Science, Services and Agents on the World*

Wide Web, *44*, 3–36. Descargado de http://www.sciencedirect.com/science/article/pii/S1570826817300276 doi: https://doi.org/10.1016/j.websem.2017.05.005

Kitchenham, B., Pickard, L., y Pfleeger, S. L. (1995). Case studies for method and tool evaluation. *IEEE software*, *12*(4), 52–62.

Klyne, G., y Caroll, J. J. (2004). *Resource Description Framework (RDF): Concepts and Abstract Syntax.* Descargado 2017/12/28, de http://www.w3.org/TR/rdf-concepts/

K. Munir, M. O., y McClatchey, R. (2012). Ontology-driven relational query formulation using the semantic and assertional capabilities of OWL-DL. *Knowledge-based Systems*, *35*, 144–159.

Konstantinou, N., Houssos, N., y Manta, A. (2014). Exposing Bibliographic Information as Linked Open Data Using Standards-based Mappings: Methodology and Results. *Procedia - Social and Behavioral Sciences*, *147*(1), 260–267. Descargado de http://www.sciencedirect.com/science/article/pii/S1877042814040889 doi: 10.1016/j.sbspro.2014.07.169

Lacasta, J., Nogueras-Iso, J., Falquet, G., Teller, J., y Zarazaga-soria, F. J. (2013). Design and evaluation of a semantic enrichment process for bibliographic databases. *Data & Knowledge Engineering*, *88*(1), 94–107. Descargado de http://dx.doi.org/10.1016/j.datak.2013.10.001 doi: 10.1016/j.datak.2013.10.001

Leavitt, N. (2010). Will NoSQL Databases Live Up to Their Promise? *Computer*, *43*(2). Descargado de http://ieeexplore.ieee.org/abstract/document/5410700/ doi: https://doi.org/10.1109/MC.2010.58

Leiva-Mederos, A., Senso, J. A., Domínguez-Velasco, S., y Hípola, P. (2013). AUTHORIS: a tool for authority control in the semantic web. *Library Hi Tech*, *31*(3), 536–553. doi: 10.1108/LHT-12-20112-0135

Lenzerini, M. (2002). Data integration: A theoretical perspective. En *Proceedings of the twenty-first acm sigmod-sigact-sigart symposium on principles of database systems* (pp. 233–246). New York, NY, USA: ACM. Descargado de http://doi.acm.org/10.1145/543613.543644 doi: 10.1145/543613.543644

Lethbridge, T. C., Sim, S. E., y Singer, J. (2005, 01 de Jul). Studying software engineers: Data collection techniques for software field studies. *Empirical Software Engineering*, *10*(3), 311–341. Descargado de https://doi.org/10.1007/s10664-005-1290-x doi: 10.1007/s10664-005-1290-x

Lubetzky, S., y Hayes, R. M. (1969). *The principles of cataloging: Report*. Institute of Library Research, University of California.

Macina, A., Montagnat, J., y Corby, O. (2016). A SPARQL distributed query processing engine addressing both vertical and horizontal data partitions. En *Bda 2016-32ème conférence sur la gestion de données-principes, technologies et applications*. Poitiers, France. Descargado de https://hal.archives-ouvertes.fr/hal-01404165/file/macina{_}montagnat{_}corby{_}bda2016.pdf

Maier, D. (1983). *The theory of relational databases*. Rockville, MD: Computer science press Rockville.

Makki, J., Alquier, A.-M., y Prince, V. (2008). Semi Automatic Ontology Instantiation in the domain of Risk Management. En Z. Shi, E. Mercier-Laurent, y D. Leake (Eds.), *Intelligent information processing iv* (pp. 254–265). Boston, MA: Springer US.

March, S. T., y Smith, G. F. (1995). Design and Natural Science Research on Information Technology. *Decision Support Systems*, *15*(4), 251–266. Descargado de http://dx.doi.org/10.1016/0167-9236(94)00041-2 doi: 10.1016/0167-9236(94)00041-2

Masinter, L., Berners-Lee, T., y Fielding, R. T. (2005). *Uniform Resource Identifier*. Descargado 12/28/2017, de http://tools.ietf.org/html/rfc3986

McGinnes, S., y Kapros, E. (2015). Conceptual independence: A design principle for the construction of adaptive information systems. *Information Systems*, *47*, 33–50. doi: 10.1016/j.is.2014.06.001

Michel, F. (2017). *Integrating Heterogeneous Data Sources in the Web of Data* (Tesis Doctoral). Côte DÁzur.

Moniruzzaman, A. B. M., y Hossain, S. A. (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *International Journal of Database Theory and Application*, *6*(4), 1–14.

Motik, B., Horrocks, I., y Sattler, U. (2009). Bridging the gap between owl and relational databases. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7*(2), 74–89.

Nachouki, G., y Quafafou, M. (2011). MashUp web data sources and services based on semantic queries. *Information Systems*, *36*(2), 151–173. Descargado de http://www.elsevier.com/locate/infosys doi: 10.1016/j.is.2010.08.001

Offermann, P., Blom, S., Schönherr, M., y Bub, U. (2010). Artifact Types in Information Systems Design Science – A Literature Review. En R. Winter, Leon Zhao J., y S. Aier (Eds.), *Global perspectives on design science research* (pp. 77–92). St. Gallen: Springer-Verlag Berlin Heidelberg. Descargado de https://link.springer.com/book/10.1007/978-3-642-13335-0{#}toc doi: https://doi.org/10.1007/978-3-642-13335-0

Online Computer Library Center Inc. (2017). *VIAF*. Descargado 2017-12-15, de http://www.oclc.org/viaf.en.html

ORCID. (2017). *ORCID | Connecting Research and Researchers*. Descargado 2017-12-15, de https://orcid.org/

Pautasso, C. (2014). RESTful Web Services: Principles, Patterns, Emerging Technologies. En A. Bouguettaya, Q. Z. Sheng, y F. Daniel (Eds.), *Web services foundations* (pp. 31–51). New York, NY: Springer New York. Descargado de https://doi.org/10.1007/978-1-4614-7518-7{_}2 doi: 10.1007/978-1-4614-7518-7_2

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., y Bragge, J. (2006). The Design Science Research Process: A Model for Producing and Presenting Information Systems Research. En *Proceedings of design research in information systems and technology desrist'06* (Vol. 24, pp. 83–106). doi: 10.2753/MIS0742-1222240302

Petersen, K., y Wohlin, C. (2009). Context in Industrial Software Engineering Research. En *Proceedings of the 2009 3rd international symposium on empirical software engineering and measurement* (pp. 401–404). Washington, DC, USA: IEEE Computer Society. Descargado de http://dx.doi.org/10.1109/ESEM.2009.5316010 doi: 10.1109/ESEM.2009.5316010

Pokorny, J. (2013). NoSQL databases: a step to database scalability in web environment. *International Journal of Web Information Systems*, *9*(1), 69–82. Descargado de http://dl.acm.org/citation.cfm?doid=2095536.2095583 doi: 10.1108/17440081311316398

Real Academia Española. (2014). *Diccionario de la lengua española :[Edición del Tricentenario]*. Descargado 2017-12-13, de http://dle.rae.es/

Runeson, P., Host, M., Rainer, A., y Regnell, B. (2012). *Case Study Research in Software Engineering* (1st ed.). Wiley Publishing. Descargado de http://www.wiley.com/WileyCDA/WileyTitle/productCd-1118104358.html{%}5Cnhttp://doi.wiley.com/10.1002/9781118181034 doi: 10.1002/9781118181034

Sandberg, J., y Jin, Q. (2016). How should Catalogers Provide Authority Control for Journal Article Authors? Name Identifiers in the Linked Data World. *Cataloging & Classification Quarterly*, *54*(8), 1–16. Descargado de http://eprints.rclis.org/30155/ doi: http://dx.doi.org/10.1080/01639374.2016.1238429

Schwarte, A., Haase, P., Hose, K., Schenkel, R., y Schmidt, M. (2011). FedX: Optimization Techniques for Federated Query Processing on Linked Data. En *Proceedings of the 10th international conference on the semantic web - volume part i* (pp. 601–616). Berlin, Heidelberg: Springer-Verlag. Descargado de http://dl.acm.org/citation.cfm?id=2063016.2063055

Seaman, C. B. (1999, Jul). Qualitative methods in empirical studies of software engineering. *IEEE Transactions on Software Engineering*, *25*(4), 557-572. doi: 10.1109/32.799955

Shanmugasundaram, J., Tufte, K., He, G., Zhang, C., DeWitt, D. J., Naughton, J. F., ... Naughton, J. F. (1999). Relational databases for querying XML documents: Limitations and opportunities. En M. P. Atkinson, M. E. Orlowska, P. Valduriez, S. B. Zdonick, y M. L. Brodie (Eds.), *Proceedings of the 25th vldb conference,* (pp. 302–314). San Francisco, CA: Morgan Kaufmann Publishers Inc. doi: 10.1016/j.acalib.2005.12.008

Sheth, A. P., y Larson, J. A. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, *22*(3), 183–236. Descargado de http://portal.acm.org/citation.cfm?doid=96602.96604 doi: 10.1145/96602.96604

Simon, H. A. (1996). *The Sciences of the Artificial* (3rd ed.). Cambridge, MA, USA: MIT Press.

Spanos, D.-E., Stavrou, P., y Mitrou, N. (2012). Bringing relational databases into the Semantic Web: A survey. *Semantic Web*, *3*(2), 169–209. Descargado de http://www.semantic-web-journal.net/sites/default/files/swj121_1.pdf doi: 10.3233/SW-2011-0055

Staab, S., y Studer, R. (Eds.). (2009). *Handbook on Ontologies* (2.ª ed.). New York, NY: Springer-Verlag Berlin Heidelberg. doi: 10.1007/978-3-540-92673-3

Stake, R. E. (1995). *The art of case study research*. SAGE Publications.

Studer, R., Richards Benjamins, V., y Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, *25*(1-2), 161–197. doi: 10.1016/S0169-023X(97)00056-6

Sulé, A., Centelles, M., Franganillo, J., y Gascón, J. (2016). Aplicación del modelo de datos RDF en las colecciones digitales de bibliotecas, archivos y museos de España. *Revista española de Documentación Científica*, *39*(1), 121–139. Descargado de http://redc.revistas.csic.es/index.php/redc/article/view/924/1340 doi: 10.3989/redc.2016.1.1268

Tabares-Martín, L., Fernández-Peña, F. O., y Leiva-Mederos, A. (2016). AUCTORITAS : A Semantic Web-based tool for Authority Control. En Y. Hidalgo-Delgado y A. A. Leiva-Mederos (Eds.), *2nd international workshop of*

semantic web (pp. 11–22). Havana: cewr-ws. Descargado de http://ceur-ws.org/Vol-1797/paper2.pdf

Tabares-Martín, L., Fernández-Peña, F. O., Leiva-Mederos, A., Goovaerts, M., Calzadilla-Reyes, D., y Ruano-Alvarez, W. A. (2015). Software Applications Ecosystem for Authority Control. En *Metadata and semantics research conference* (pp. 214–224). Garoufallou, Emanouel. doi: 10.1007/978-3-319-24129-6

Tabares-Martín, L., Fernández-Peña, F. O., Leiva-Mederos, A., y Nummenmaa, J. (2016). An empirical performance evaluation of a semantic-based data retrieving process from RDBs & RDF data storages. *MASKANA*, *7*, 23–34. Descargado de http://dspace.ucuenca.edu.ec/jspui/handle/123456789/26335

Tillett, B. B. (2009). Authority Control : State of the Art and New Perspectives. *Cataloging & Classification Quarterly*, *38*(3/4), 23–41. doi: 10.1300/J104v38n03_04

Vavliakis, K. N., Grollios, T. K., y Mitkas, P. A. (2013). RDOTE - Publishing relational databases into the semantic web. *Journal of Systems and Software*, *86*(1), 89–99. doi: 10.1016/j.jss.2012.07.018

Verborgh, R., Sande, M. V., Hartig, O., Herwegen, J. V., Vocht, L. D., Meester, B. D., ... Colpaert, P. (2016). Triple Pattern Fragments: A low-cost knowledge graph interface for the Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, *37-38*, 184–206. Descargado de http://www.sciencedirect.com/science/article/pii/S1570826816000214 doi: https://doi.org/10.1016/j.websem.2016.03.003

Viljanen, K., Tuominen, J., y Hyvönen, E. (2008). Publishing and using ontologies as mashup services. En *Proceedings of the 4th workshop on scripting for the semantic web (sfsw2008)*. 5th European Semantic Web Conference.

W3C SPARQL Working Group. (2013). *SPARQL 1.1 Overview*. Descargado 5/3/2018, de https://www.w3.org/TR/sparql11-overview/

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., y Wesslén, A. (2012). *Experimentation in software engineering* (Vol. 9783642290). 3642290434, 9783642290435. doi: 10.1007/978-3-642-29044-2

Zaki, N., Tennakoon, C., y Al-Ashwal, H. (2017). Knowledge graph construction and search for biological databases. En *International conference on research and innovation in information systems (icriis)*. Langkawi, Malaysia: IEEE Computer Society. Descargado de http://ieeexplore.ieee.org/abstract/document/8002465/ doi: https://doi.org/10.1109/ICRIIS.2017.8002465

Bibliografía

- Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., & Shadbolt, N. R. (2003). Automatic ontology-based knowledge extraction from Web documents. IEEE Intelligent Systems, 18(1), 14–21. https://doi.org/10.1109/MIS.2003.1179189
- Bernstein, A., Hendler, J., & Noy, N. (2016). A new look at the semantic web. Communications of the ACM, 59(9), 35–37. https://doi.org/10.1145/2890489
- Botoeva, E., Calvanese, D., Cogrel, B., Rezk, M., & Xiao, G. (2016). OBDA beyond relational DBs: A study for MongoDB. In CEUR Workshop Proceedings (Vol. 1577).
- Fernández-López, M. (1999). Overview Of Methodologies For Building Ontologies. In V. R. Benjamins, B. Chandrasekaran, A. Gómez-Pérez, N. Guarino, & M. Uschold (Eds.), Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5) (Vol. 18, pp. 1–13). Estocolmo, Suecia: CEUR-WS.
- Fernández-Peña, F., Acosta-Sánchez, R., & Ponce-Toste, Y. (2015). ViewOnto: modelo conceptual para la generación automática de vistas de datos. Ciencias de La Información, 46(1), 19–25.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75–105.
- Jannach, D., Shchekotykhin, K., & Friedrich, G. (2009). Automated ontology instantiation from tabular web sources—The AllRight system. Web Semantics: Science, Services and Agents on the World Wide Web, 7(3), 136–153. https://doi.org/https://doi.org/10.1016/j.websem.2009.04.002
- Kharlamov, E., Brandt, S., Jimenez-Ruiz, E., Kotidis, Y., Lamparter, S., Mailis, T., … Moeller, R. (2016). Ontology-Based Integration of Streaming and Static Relational Data with Optique. In Proceedings of the 2016 International Conference on Management of Data (pp. 2109–2112). New York, NY, USA: ACM. https://doi.org/10.1145/2882903.2899385
- Kharlamov, E., Hovland, D., Skjæveland, M. G., Bilidas, D., Jiménez-Ruiz, E., Xiao, G., … Waaler, A. (2017). Ontology Based Data Access in Statoil. Web Semantics: Science, Services and Agents on the World Wide Web, 44, 3–36. https://doi.org/https://doi.org/10.1016/j.websem.2017.05.005
- Runeson, P., Host, M., Rainer, A., & Regnell, B. (2012). Case Study Research in Software Engineering. John Wiley & Sons, Inc (1st ed.). Wiley Publishing. https://doi.org/10.1002/9781118181034

Anexo A

Glosario de términos

Requisitos funcionales para datos de autoridad: Modelo conceptual con el objetivo de proveer un marco de trabajo para el análisis de los datos de autoridad requeridos para soportar el Control de Autoridades y el intercambio internacional de datos de autoridad.

Federación Internacional de Asociaciones de Bibliotecarios y Bibliotecas: Organización mundial creada para proporcionar a los bibliotecarios un foro para intercambiar ideas, promoviendo la cooperación, la investigación y el desarrollo internacionales en todos los campos relacionados con la actividad bibliotecaria y la bibliotecología.