

**Autor:** Ing. Marco Javier Scalone Romano <mscalone@fing.edu.uy>.

**Título de tesis:** Uso de Datos Enlazados para la publicación e integración de datos de índole académico.

**Tutora:** Dra. Ing. Regina Motz <rmotz@fing.edu.uy>, Universidad de la República / Facultad de Ingeniería, Instituto de Computación.

**Fecha de defensa:** 27 de febrero de 2019.

**Resumen:** Hoy en día existe una gran cantidad de fuentes de información bibliográfica y de repositorios institucionales abiertos en línea. Estas fuentes, independientes, heterogéneas y distribuidas, suelen representar sus datos de diferente forma y brindar acceso a través de distintos mecanismos o protocolos. Además existe el grave problema de que no es costumbre identificar de forma unívoca a los autores de las publicaciones, a pesar que esto se ha comenzado a solucionar por el uso de ORCID, su utilización no es aún extendida fuera de los ámbitos de algunos servicios de publicación y no es para nada utilizado todavía en los ámbitos educativos. El mayor problema ocurre al integrar datos de fuentes de publicaciones científicas con fuentes como páginas web personales o institucionales o espacios de creaciones de materiales donde acostumbran trabajar los docentes investigadores. Es en este escenario de docentes-investigadores que este trabajo estudia el ciclo de vida de la publicación de Linked Data (Datos Enlazados) como una forma de resolver el problema de integración de datos de publicaciones científicas. Este trabajo presenta un análisis de los conceptos de la web semántica aplicados a la publicación de Datos Enlazados y una revisión de las metodologías, recomendaciones y buenas prácticas existentes para la publicación de Datos Enlazados en la web. Estas guías y recomendaciones son utilizadas como base para el análisis de dos casos de estudio que se presentan, ambos de características diferentes, como lo son los libros de texto creados en la plataforma CNX.org, y la publicación, integración y análisis de las publicaciones científicas producidas por los docentes del Instituto de Computación de la Facultad de Ingeniería (UdelaR). En este último caso se publicaron como Datos Enlazados la lista de docentes publicada en el sitio web de la institución y las bases bibliográficas disponibles en el sitio web de FIng y en DBLP. Se diseñaron y ejecutaron procesos de detección de enlaces y resolución de identidad entre las tres fuentes y se presenta a la vez un estudio analítico a partir del uso de los Datos Enlazados.

**Palabras clave:** Datos Enlazados, Web semántica, Integración de datos, Resolución de identidad, Bibliometría.



**PEDECIBA Informática**  
Universidad de la República  
Uruguay

# TESIS DE MAESTRÍA EN INFORMÁTICA

---

**Uso de Datos Enlazados para la  
publicación e integración de datos  
de índole académico.**

Ing. Marco Scalone

---

**Director Académico y Director de Tesis:**

Dra. Ing. Regina Motz  
Instituto de Computación  
Universidad de la República  
Uruguay

2018

## Resumen

Hoy en día existe una gran cantidad de fuentes de información bibliográfica y de repositorios institucionales abiertos en línea. Estas fuentes, independientes, heterogéneas y distribuidas, suelen representar sus datos de diferente forma y brindar acceso a través de distintos mecanismos o protocolos. Además existe el grave problema de que no es costumbre identificar de forma unívoca a los autores de las publicaciones, a pesar que esto se ha comenzado a solucionar por el uso de ORCID, su utilización no es aún extendida fuera de los ámbitos de algunos servicios de publicación y no es para nada utilizado todavía en los ámbitos educativos. El mayor problema ocurre al integrar datos de fuentes de publicaciones científicas con fuentes como páginas web personales o institucionales o espacios de creaciones de materiales donde acostumbran trabajar los docentes-investigadores. Es en este escenario de docentes-investigadores que este trabajo estudia el ciclo de vida de la publicación de Linked Data (Datos Enlazados) como una forma de resolver el problema de integración de datos de publicaciones científicas. Este trabajo presenta un análisis de los conceptos de la web semántica aplicados a la publicación de Datos Enlazados y una revisión de las metodologías, recomendaciones y buenas prácticas existentes para la publicación de Datos Enlazados en la web. Estas guías y recomendaciones son utilizadas como base para el análisis de dos casos de estudio que se presentan, ambos de características diferentes, como lo son los libros de texto creados en la plataforma CNX.org, y la publicación, integración y análisis de las publicaciones científicas producidas por los docentes del Instituto de Computación de la Facultad de Ingeniería (UdelaR). En este último caso se publicaron como Datos Enlazados la lista de docentes publicada en el sitio web de la institución y las bases bibliográficas disponibles en el sitio web de FIng y en DBLP. Se diseñaron y ejecutaron procesos de detección de enlaces y resolución de identidad entre las tres fuentes y se presenta a la vez un estudio analítico a partir del uso de los Datos Enlazados.

**Palabras clave:** Datos Enlazados, Web semántica, Integración de datos, Resolución de identidad, Bibliometría.

# Índice

<b>1. Introducción</b>	<b>1</b>
1.1. Motivación . . . . .	1
1.2. Objetivos . . . . .	3
1.3. Organización del documento . . . . .	3
<b>2. Marco conceptual</b>	<b>4</b>
2.1. Web Semántica . . . . .	4
2.1.1. Introducción . . . . .	4
2.1.2. Arquitectura . . . . .	5
2.1.3. Ontologías . . . . .	7
2.1.4. Lenguajes de consulta . . . . .	11
2.2. Datos Enlazados . . . . .	13
2.2.1. Principios básicos de Datos Enlazados . . . . .	14
2.2.2. Vocabularios para describir los datos . . . . .	16
2.2.3. Ciclo de vida . . . . .	17
2.2.4. Buenas prácticas . . . . .	21
2.2.5. Patrones de publicación . . . . .	22
2.2.6. Estado actual de la Web de Datos . . . . .	24
<b>3. Aplicación</b>	<b>27</b>
3.1. CNX.org . . . . .	28
3.1.1. Especificación . . . . .	29
3.1.2. Modelado . . . . .	37
3.1.3. Generación de Datos . . . . .	37
3.1.4. Publicación . . . . .	44
3.2. Producción bibliográfica del InCo . . . . .	46
3.2.1. Especificación . . . . .	47
3.2.2. Modelado . . . . .	51
3.2.3. Generación . . . . .	53
3.2.4. Publicación . . . . .	63

3.2.5. Análisis de resultados . . . . .	66
<b>4. Conclusiones y trabajos futuros</b>	<b>72</b>
4.1. Conclusiones . . . . .	72
4.1.1. Trabajos similares . . . . .	73
4.2. Trabajo Futuro . . . . .	74
<b>Bibliografía</b>	<b>79</b>
<b>Abreviaciones</b>	<b>81</b>
<b>Nomenclatura</b>	<b>84</b>
<b>Apéndices</b>	
A. Caso CNX	85
B. Caso InCo	90

# Lista de Figuras

2.1.	Diagrama de pila de tecnologías y estándares de la Web Semántica .	6
2.2.	Ejemplo de grfo RDF simple representado gráficamente . . . . .	9
2.3.	Ejemplo de grfo RDF simple representado gráficamente . . . . .	10
2.4.	Ciclo de vida planteado en LOD2 Stack . . . . .	20
2.5.	Patrones de publicación según la naturaleza de los datos [24] . . . .	23
2.6.	Grafo con la estructura general y la categorización de los juegos de datos por dominio temático. El tamaño de lo círculos representa el grado de referencias que recibe desde otros juegos de datos. . . . .	25
3.1.	Ubicación del caso de estudio CNX.org en los Patrones de publicación	35
3.2.	Componentes de OAI2LOD Server [31]. . . . .	38
3.3.	Diagrama de clases simplificado donde se muestran los componentes principales de OAI2LOD Server así como las clases que fueron creadas o modificadas (en color verde) . . . . .	41
3.4.	Ejemplo de visualización de una colección desde un navegador web.	45
3.5.	Ubicación del escenario teniendo en cuenta la naturaleza de los datos	48
3.6.	Representación gráfica del vocabulario Docentes. Las propiedades en color verde fueron definidas en el vocabulario, mientras que las azules se reutilizaron de vocabularios existentes. . . . .	52
3.7.	Representación gráfica del vocabulario bibtex utilizado: <a href="http://purl.org/net/nknouf/ns/bibtex#">http://purl.org/net/nknouf/ns/bibtex#</a> . . . . .	53
3.8.	Diagrama de extracción transformación y carga de RDF . . . . .	55
3.9.	Ejemplo de revisión y validación de enlaces detectados entre publicaciones de <i>Biblio FIng</i> y DBLP utilizando la herramienta <i>Silk Workbench</i> . . . . .	56
3.10.	Enlaces entre los tipos de entidades de las distintas fuentes. Se muestra entre paréntesis la cantidad de tripletas. . . . .	57
3.11.	Diagrama del ciclo de vida del procesamiento de enlaces de DBpedia [34] . . . . .	58
3.12.	Regla de comparación definida en <i>Silk Workbench</i> entre publicaciones de <i>Biblio FIng</i> y DBLP . . . . .	59

3.13. Diagrama de regla de comparación entre docentes del InCo y coautores de DBLP . . . . .	61
3.14. Cantidad de docentes del InCo identificados en DBLP distribuidos por grado. . . . .	62
3.15. Ejemplo de visualización de los datos de un docente accediendo desde un navegador. . . . .	65
3.16. Promedio de publicaciones por docente y tipo de publicación. . . . .	67
3.17. Promedio de coautorías dentro del InCo por docente y tipo de publicación. . . . .	67
3.18. Histograma de cantidad de publicaciones por grado en intervalos de 5 publicaciones. . . . .	67
3.19. Desviación estándar del promedio de coautorías entre docentes del InCo por grado y tipo de publicación. . . . .	67
3.20. Distribución de publicaciones por tipo (izquierda) y por tipo y grado (derecha) . . . . .	69
3.21. Distribución publicaciones del InCo por año y tipo . . . . .	69
3.22. Distribución de todas las publicaciones de DBLP por año y tipo . . . . .	70

# Lista de Tablas

2.1. Resumen cumplimiento de buenas prácticas generado a partir de los datos publicados en [28] . . . . .	26
3.1. Lista de prefijos de URIs utilizados a lo largo del documento . . . . .	28
3.2. Metadatos soportados en el formato MDML . . . . .	31
3.3. Mecanismos y nivel de acceso a metadatos . . . . .	34
3.4. Resumen de los distintos métodos de acceso . . . . .	35
3.5. Cantidad de entidades que usan los distintos sabores de licencia <i>Creative Commons</i> en la plataforma . . . . .	36
3.6. Cantidad de documentos en cada tema principal con su respectiva categoría DBpedia vinculada . . . . .	42
3.7. Autores identificados en recursos de tipo Persona en DBpedia y la cantidad de publicaciones de CNX en los que figura como autor. . . . .	43
3.8. Mapeo de los datos de Docentes disponibles propiedades de distintos vocabularios. . . . .	51
3.9. Población de las distintas entidades de la fuente <i>Biblio FIng</i> asociada a la cantidad enlaces detectados desde la fuente <i>Docentes InCo</i> y hacia la fuente DBLP. . . . .	60
3.10. Cantidad de docentes del InCo identificados en DBLP distribuidos por grado. . . . .	61
3.11. Ejemplos de ambigüedades detectadas durante la generación de enlaces entre docentes del InCo y autores de DBLP . . . . .	63
3.12. Ejemplo de redirección para la dereferenciación de URIs. . . . .	64
3.13. Top 20 de conferencias y revistas con más publicaciones en el período 2006-2016 . . . . .	71



# Capítulo 1

## Introducción

### 1.1. Motivación

La educación es una de las áreas en las que Internet ha tenido un alto impacto en los últimos años. Las tecnologías web han mejorado el soporte a muchas tareas relacionadas con la enseñanza, el aprendizaje y la investigación. Una de ellas es la publicación de libros de texto de forma electrónica. En los últimos años se han popularizado varias plataformas para la creación y publicación de este tipo de contenidos que facilitan el trabajo de creación, edición y revisión por personas geográficamente distribuidas. Cada vez son más los recursos disponibles en línea para analizar y combinar. El problema ya no es acceder a la información, sino poder seleccionar la más relevante en un tiempo razonable con el menor esfuerzo posible.

De igual modo en el ámbito de la investigación el número de publicaciones ha ido en aumento de forma sostenida. Existe una gran cantidad de fuentes de información bibliográfica en línea, tales como los sitios web de las editoriales, revistas o conferencias científicas. A su vez se ha popularizado el uso de repositorios institucionales abiertos, donde se almacenan trabajos académicos como tesis de maestría o doctorado. Esto ha generado un interés de la comunidad académica por gestionar y analizar la información bibliográfica. Este interés ha dado lugar al surgimiento de diversas iniciativas que intentan facilitar la búsqueda y el acceso a dicha infor-

mación tales como *Research Gate*<sup>1</sup>, *Semantic Scholar*<sup>2</sup> o *Google Scholar*<sup>3</sup>, entre otras.

A su vez esta abundancia de información ha promovido el desarrollo en áreas de investigación tales como *Learning Analytics*, *Educational Data Mining* o la *Bibliometría*, que investigan y analizan aspectos relacionados con el aprendizaje y la actividad científica. El progreso en estas áreas depende tanto de la cantidad y calidad de los datos disponibles, como de la capacidad para interpretar claramente el significado de los mismos. Sin embargo, si bien la información en línea es abundante, ésta se encuentra distribuida en fuentes heterogéneas e independientes, accesibles a través de distintos mecanismos o protocolos y representada en diferentes formatos. Muchas veces las fuentes utilizan distintos modelos e identificadores para representar las mismas entidades, lo que genera ambigüedades y problemas de resolución de identidad. Esto dificulta enormemente la integración de las fuentes y el análisis de los datos.

En paralelo los principios de los Datos Enlazados (*Linked Data* en inglés) se han ido consolidando como una forma de exponer datos en la web con el potencial de solucionar varias de las limitaciones anteriormente mencionadas. Los Datos Enlazados son una forma de Web Semántica que se compone de un conjunto de buenas prácticas para publicar y enlazar datos estructurados en la web. Los datos y su significado se definen explícitamente de forma tal que no requiera de un humano para ser interpretados. Para representar y consultar los datos se apoya en establecidos por la *World Wide Web Consortium (W3C)* tales como *Resource Description Framework (RDF)* y *SPARQL Protocol and RDF Query Language (SPARQL)*. Esto hace que los Datos Enlazados constituyan una capa ideal de gestión de datos para el escenario aquí planteado.

Con el fin de explorar y analizar estos temas se definieron dos casos de uso. El primero es el análisis de la plataforma de creación y publicación de libros de texto CNX.org. Fue una de las plataformas candidatas para ser usada en el proyecto LATIn<sup>4</sup> y cuenta con más de mil colecciones (libros, cursos, etc.) que utilizan y combinan más de 15000 objetos de aprendizaje (llamados módulos). El segundo caso corresponde al análisis de la producción bibliográfica de los docentes del *Instituto de Computación (InCo)* de la *Facultad de Ingeniería (FIng)* de la *Universidad de la República (UdelaR)*. Para esto se debieron analizar e integrar varias fuentes de datos con el fin de poder obtener una visión integrada de los datos.

---

<sup>1</sup><https://www.researchgate.net/>

<sup>2</sup><https://www.semanticscholar.org/>

<sup>3</sup><https://scholar.google.com.uy/>

<sup>4</sup><http://proeva.udelar.edu.uy/institucional/proyectos/latin/>

## 1.2. Objetivos

El objetivo general de este estudio es investigar las etapas involucradas en el proceso de publicación e integración de datos, usando Datos Enlazados. El análisis se enmarca en el contexto de publicaciones de índole académico, tales como libros de texto de la plataforma CNX.org y trabajos de investigación de docentes del InCo. De aquí se desprenden algunos objetivos específicos que sirvieron de guía para el desarrollo de la tesis:

- Identificar las técnicas, procesos y herramientas relacionadas con la publicación y uso de Datos Enlazados en el contexto de publicaciones académicas.
- Analizar el funcionamiento de la plataforma CNX, en particular la forma de representación y acceso de los datos.
- Analizar las fuentes de datos de información académica disponibles relacionadas con los docentes del InCo.
- Teniendo en cuenta lo analizado en los puntos anteriores proponer un posible mecanismo para la publicación de dichos datos como Datos Enlazados.
- Demostrar la viabilidad de la propuesta implementando un prototipo que realice la publicación de dichos datos y posibilite la integración con otras fuentes existentes.

## 1.3. Organización del documento

El documento se organiza de la siguiente manera: la sección 2 presenta los conceptos más relevantes que se manejan a lo largo del documento, así como el contexto tecnológico y metodológico en que se desarrolla el proyecto. En la sección 3 se presenta la propuesta para el abordaje de los casos de uso siguiendo el marco conceptual y metodológico descrito en la sección 2. Finalmente, la sección 4 presenta las conclusiones del trabajo realizado y los resultados obtenidos, así como posibles trabajos a futuro.

# Capítulo 2

## Marco conceptual

### 2.1. Web Semántica

#### 2.1.1. Introducción

La web ha cambiado radicalmente la forma en que compartimos el conocimiento, facilitando la publicación y acceso a una enorme cantidad y diversidad de información. La naturaleza genérica, abierta y extensible la ha convertido en uno de los repositorios más grandes de información en prácticamente todas las áreas de conocimiento. Esta abundancia de información hace que sea cada vez más difícil filtrar lo relevante entre la inmensa cantidad de datos acumulados a lo largo del tiempo. Tal es el punto que los motores de búsqueda han jugado un papel fundamental en el desarrollo de la web. Su evolución a lo largo del tiempo los han transformado en factor clave para el acceso a la información. Sin embargo los desafíos originales de obtener buena **precisión** y **sensibilidad** en los resultados de búsqueda se mantienen. Sigue siendo difícil filtrar el contenido obsoleto o ambiguo, y no es posible controlar la granularidad de los resultados. Esto sucede porque la limitación de base sigue siendo la misma: *interpretar el significado del contenido*.

La idea de extender las capacidades de la web publicando datos estructurados no es nueva y se remite prácticamente a los orígenes de su propuesta [1]. Desde sus inicios la web fue pensada para que la "evolución de los objetos pase de ser legible por humanos a contener más información semántica orientada a las computadoras"[2]. Sin embargo la gran mayoría de la información aun se encuentra en lenguaje natural o estructuras de datos complejas como audio o video. Esto hace

que los agentes de software que la procesan se vean obligados a utilizar heurísticas más complejas y potencialmente inexactas para interpretar de forma automática el conocimiento subyacente. El concepto de Web Semántica surge como una respuesta a esas limitaciones.

En una de las primeras publicaciones presentadas al respecto Tim Berners-Lee plantea la visión de una web enriquecida, donde el significado de los distintos recursos disponibles se presente de forma explícita. Esto permitiría que las máquinas puedan entender y procesar la información de forma precisa y eficiente, dando soporte a nuevos y mejores servicios automatizados para las personas [3]. Esto no solo permite el descubrimiento e interpretación automática de los datos, sino que además da la posibilidad de inferir nuevo conocimiento a partir del disponible.

La Web Semántica se posiciona como una extensión de la web actual, y al igual que en esta última los lineamientos y tecnologías que la componen son regulados por la W3C. La W3C es un consorcio internacional integrado por una gran cantidad de organizaciones de todo el mundo<sup>5</sup>. Su misión es generar estándares y recomendaciones que aseguren el crecimiento y evolución de la web a largo plazo.

Se puede decir que la Web Semántica aún se encuentra en una etapa temprana de desarrollo, sin embargo las tecnologías y lineamientos que le dan soporte ya están siendo adoptadas en diversos ámbitos como la industria, el gobierno o la academia. Los escenarios de aplicaciones varían desde la integración de datos y servicios hasta el descubrimiento y recomendación de contenido. Incluso los motores de búsqueda han incorporado estas tecnologías para proporcionar resultados de búsqueda más completos y específicos<sup>6</sup>. En [4] se puede encontrar una amplia lista de casos de estudio de aplicación de estas tecnologías.

### 2.1.2. Arquitectura

La arquitectura de la Web Semántica suele describirse por medio de un diagrama de pila de tecnologías conocido como *Semantic Web Stack* (o *Layercake Diagram*). Dicho diagrama es una ilustración de capas organizadas jerárquicamente, donde se muestran de forma general y simplificada los estándares y tecnologías más importantes que le dan soporte. Cada capa utiliza las capacidades de las capas inferiores.

---

<sup>5</sup><https://www.w3.org/Consortium/Member/List>

<sup>6</sup>Google Knowledge Graph: <https://developers.google.com/knowledge-graph/how-tos/search-widget>

La versión original fue presentada por Tim Berners-Lee<sup>7</sup> y ha evolucionado a lo largo del tiempo conforme las tecnologías de cada capa se fueron estandarizando o consensuando [5]. Al día de hoy existen diferentes versiones de este diagrama, algunos con mayor nivel de granularidad. En la figura 2.1 se presenta una de las versiones más referenciadas de este diagrama presentado por la W3C en 2007<sup>8</sup>.

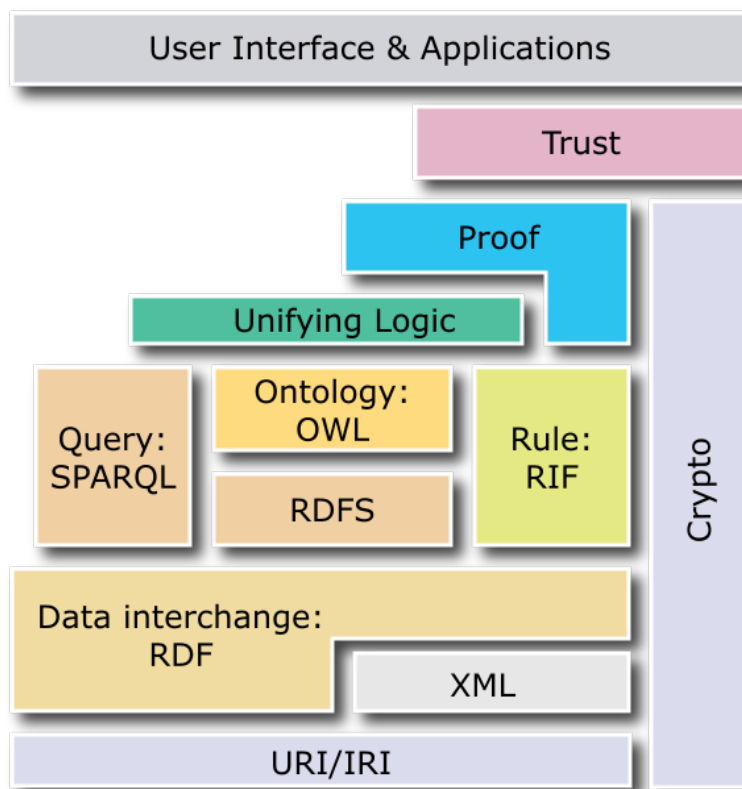


Figura 2.1: Diagrama de pila de tecnologías y estándares de la Web Semántica

Las capas inferiores especifican conceptos de más bajo nivel relacionados con el transporte, la identificación y la serialización de los datos (**URI/UNICODE/XML**). En paralelo, y a lo largo de varias capas, se utilizan servicios de criptografía para asegurar la integridad y autenticidad de los datos que se intercambian.

Por encima se encuentran los estándares para definir el modelo y la semántica adicional de los datos (**RDF/RDFS/OWL**). Estas tecnologías permiten la definición de un esquema que describa los distintos conceptos de la realidad, sus relaciones

<sup>7</sup>Semantic Web - XML 2000: <https://www.w3.org/2000/Talks/1206-xml2k-tbl/slide1-0.html>

<sup>8</sup>[https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#\(24\)](https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(24))

y las restricciones que los caracterizan. También incluyen la capacidad de realizar consultas complejas sobre los datos (**SPARQL**) y especificar reglas de razonamiento (**RIF/SWRL**) que permitan inferir nuevos hechos que no están explícitos en los datos.

Finalmente las capas superiores proveen servicios de un mayor nivel semántico, que apuntan a mejorar la robustez de las aplicaciones que se desarrollen por encima. Proporcionan la capacidad de realizar demostraciones matemáticas o lógicas, sobre los razonamientos que se generan a partir de los datos. En el contexto semántico ideal, un agente de software debería ser capaz de justificar y comunicar las razones por las cuales toma una determinada decisión. Estas capas están relacionadas con la confiabilidad de la información que se maneja, así como la de los servicios y aplicaciones generadas a partir de ella.

Las capas inferiores están estandarizadas por lo que cuentan con una mayor aceptación por parte de la comunidad. Mientras que las tecnologías de las capas superiores son materia de investigación constante y aún no existe un consenso claro de cómo se implementarán. Cabe resaltar que un estándar se denomina recomendación **W3C** cuando ha alcanzado su nivel más alto de madurez, se han revisado minuciosamente y está listo para una amplia implementación. Antes de que una norma se convierta en una recomendación, pasa por las fases de ser un borrador de trabajo, recomendación candidata y propuesta de recomendación donde participan muchas personas y organizaciones vinculadas a la web[6].

### 2.1.3. Ontologías

Las ontologías constituyen un elemento fundamental para el desarrollo de las tecnologías de la Web Semántica. Son el eslabón clave para la representación de la semántica que posibilita la interpretación automática del conocimiento. El término *ontología* tiene origen en la filosofía y denota el estudio de la naturaleza del ser y la existencia. En ciencia de la computación este término refiere a la descripción formal de los conceptos y sus relaciones dentro de un dominio determinado. Una de las definiciones de ontología más citadas proviene del área de la Inteligencia Artificial y dice lo siguiente:

*Una ontología es una especificación formal y explícita de una conceptualización compartida.*[7]

El término *conceptualización* refiere a un modelo simplificado y abstracto de un dominio determinado. Tiene un propósito específico para el cual se define un voca-

ulario controlado. Una *especificación formal y explícita* indica que el modelo debe especificarse de forma detallada, utilizando un lenguaje y formalismo común que permita ser interpretado por personas y máquinas sin ambigüedades. *Compartido* implica que la ontología debe ser aceptada por un grupo de personas ya que será utilizada para representar conocimiento común. El concepto de compartido puede verse como un acuerdo dentro de un grupo de personas con el fin de facilitar la comunicación e intercambio de información.

Las ontologías son un marco conceptual que permite crear bases de datos para almacenar conocimiento, denominadas *bases de conocimiento* (*knowledge base* en inglés). Una base de conocimiento puede representarse conceptualmente como una combinación de terminologías (TBox) y aserciones (ABox). Utilizando una analogía en el ámbito de la programación, se puede pensar en un esquema XSD frente a un documento XML o una definición de clase frente a instanciación de clase. En la Web Semántica al definir una base de conocimiento, se considera importante diferenciar formalmente enunciados que se entienden como verdaderos dentro del dominio del discurso. Por ejemplo, la afirmación *cada mujer es un ser humano* así como los conceptos de *mujer* y *ser humano* se consideran parte de TBox. Mientras que aquellas afirmaciones que podrían haberse hecho en un momento dado en el tiempo, tal como, *María es una mujer* podría considerarse parte de ABox. Tales afirmaciones (ABox) hechas contra un conjunto de terminologías bien elaborado (TBox) pueden facilitar la inferencia, que puede ser útil para descubrir nuevas afirmaciones (dentro de ABox). En el ejemplo anterior se podría inferir que *María es un ser humano*. Las declaraciones en el TBox tienden a permanecer estáticas a lo largo del tiempo debido a la naturaleza de la verdad, mientras que ABox puede seguir cambiando a medida que se hacen más afirmaciones o las afirmaciones existentes se vuelven inválidas.

Las ontologías se pueden clasificar utilizando distintos criterios. Un ejemplo de ello es clasificarlas según su dependencia con el contexto de aplicación [8]:

- Ontologías genéricas o de alto nivel: describen conceptos generales o comunes a muchos dominios de aplicación, tales como el tiempo, el espacio, etc.
- Ontologías de dominio: son específicas de un dominio particular y suelen utilizar conceptos definidos en ontologías genéricas.
- Ontologías de tarea: describen características generales de artefactos o actividades relacionadas con procesos de resolución de problemas. Suelen ser válidas en distintos dominios. Por ejemplo, el concepto de objetivo, o planificación.



- Ontologías de aplicación: son usadas a bajo nivel por aplicaciones concretas. Suelen utilizar conceptos de las ontologías mencionadas anteriormente. Pueden por ejemplo definir tareas específicas en dominios particulares.

## Lenguajes para representar ontologías

Las ontologías tienen como objetivo representar el conocimiento consensuado de un modo explícito y formal, de tal manera que pueda ser compartido e interpretado por agentes de software. Por tal motivo las ontologías requieren de un lenguaje lógico y formal con el grado de expresividad suficiente para poder representar dicho conocimiento.

A lo largo del tiempo se han desarrollado diversas tecnologías que dan soporte a la especificación de ontologías. En el área de la Web Semántica algunas de las especificaciones más importantes recomendadas por la W3C para tales fines son **RDF**, **RDFS** y **OWL**.

El **Resource Description Framework (RDF)** provee un modelo de referencia para representar datos en la Web. Proporciona un estándar para la visualización e intercambio de información por agentes de software y personas.

La sintaxis abstracta de RDF codifica los datos como un conjunto de triplas, cada una compuesta por un sujeto, un predicado y un objeto. Un conjunto de triplas se denomina grafo RDF. Un grafo RDF se puede visualizar como un diagrama de nodo y de arco dirigido, en el que cada tripleta se representa como un enlace nodo-arco-nodo como se muestra en la Figura 2.2.

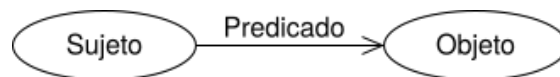


Figura 2.2: Ejemplo de grafo RDF simple representado gráficamente

El sujeto es el recurso, es decir aquello que se está describiendo. El predicado es la propiedad o relación que se desea establecer acerca del recurso. Por último, el objeto es el valor de la propiedad o el otro recurso con el que se establece la relación [9]. En la Figura 2.3 se puede observar una representación gráfica de un grafo RDF que denota que la *Persona* de nombre *Luca Scalone* nació en *Montevideo* el *12 de mayo de 2018*.

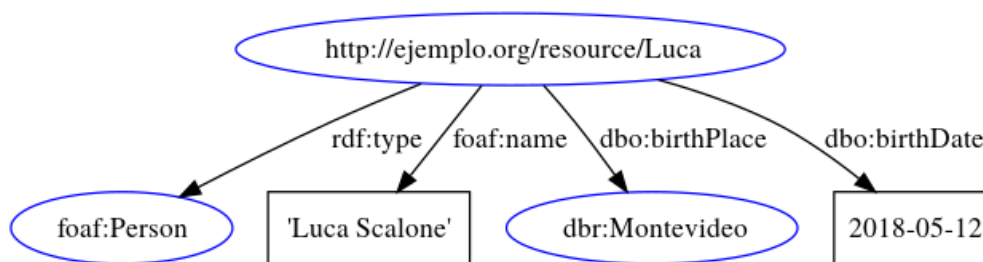


Figura 2.3: Ejemplo de grafo RDF simple representado gráficamente

Si bien **RDF** es un modelo abstracto, existen varios formatos de serialización compatibles que permiten la manipulación e intercambio de datos en RDF. Aunque estos formatos pueden describir un mismo grafo RDF de manera diferente, el conjunto de tripletas resultante es el mismo y por ende son lógicamente equivalentes. A continuación se listan algunos formatos de serialización:

- Familia de lenguajes Turtle (N-Triples, Turtle, TriG y N-Quads): Son una familia de lenguajes estrechamente relacionados que apuntan a facilitar la legibilidad por parte de las personas.
- JSON-LD: Sintaxis RDF basada en **JSON**.
- RDFa: Pensado para incrustar o embeber RDF en **HTML** y **XML**. Pensado originalmente para facilitar el procesamiento de las páginas web por parte de motores de búsqueda.
- RDF / XML : Sintaxis XML para RDF que fue originalmente propuesta por la W3C como formato de serialización.

En el Ejemplo 2.1 se presenta la serialización en formato Turtle del grafo de la Figura 2.3.

Ejemplo 2.1: Grafo RDF de la Figura 2.3 serializado en formato Turtle.

```

1  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
2  @prefix foaf: <http://xmlns.com/foaf/0.1/> .
3  @prefix eje: <http://ejemplo.org/resource/> .
4  @prefix dbo: <http://dbpedia.org/ontology/> .
5  @prefix dbr: <http://dbpedia.org/resource/> .
6
7  eje:Luca a foaf:Person ;
8           foaf:name      "Luca Scalone" ;
9           dbo:birthPlace dbr:Montevideo ;

```

El objetivo general de RDF es proporcionar un mecanismo para describir recursos que no genera ninguna asunción sobre ningún dominio de aplicación particular. No define un mecanismo para declarar estas propiedades, ni relaciones entre estas propiedades y otros recursos. Esta es la función de RDF Schema.

El **Resource Description Framework Schema (RDFS)** es una extensión de RDF que permite el refinamiento de propiedades y la definición de subclases de conceptos[10]. Esto da la posibilidad, entre otras cosas, de definir estructuras jerárquicas para representar conocimiento. RDFS es la base semántica en la mayoría de los vocabularios.

El **Web Ontology Language (OWL)**, por el contrario, posee un nivel de expresividad mayor ya que pone a disposición un vocabulario mucho más amplio para describir la realidad. Por ejemplo OWL permite utilizar operaciones de conjuntos (uniones, intersecciones) o definir restricciones de cardinalidad sobre los atributos. A modo de ejemplo, con OWL se puede expresar que la clase que representa al concepto *Persona* es disjunta de la que representa al concepto *Auto*, o que un cuarteto musical esta formado por exactamente cuatro músicos. Estas restricciones no pueden ser representadas con RDFS [11].

#### 2.1.4. Lenguajes de consulta

Contar con mecanismos de consulta que permitan recuperar un subconjunto específico de datos de forma eficiente es parte fundamental de cualquier sistema de información. Al igual que las bases de datos o XML necesitan lenguajes de consulta específicos (tales como **SQL** y **XQuery** respectivamente), RDF requiere de un mecanismo de consulta que sea compatible con su modelo de datos orientado a grafos. Aquí entra en juego **SPARQL** y es parte clave en el desarrollo de la Web Semántica que se convirtió en recomendación oficial del **W3C** en 2008 [12].

**SPARQL Protocol and RDF Query Language (SPARQL)** es un protocolo y un lenguaje de consulta a la vez. El protocolo, comúnmente denominado **SPROT** [13], describe la forma de transmitir consultas a un servicio de procesamiento de consultas **SPARQL** y devolver los resultados a la entidad que los solicitó. Los servicios que soportan este protocolo y lenguaje de consulta se denominan terminales

SPARQL (o endpoint SPARQL). Generalmente se limitan a consultar únicamente los datos almacenados en una determinada fuente, aunque en ocasiones dan la posibilidad de realizar consultas federadas. Un ejemplo destacado es el terminal SPARQL provisto por WikiData<sup>9</sup>, el cual proporciona un asistente de creación de consultas muy amigable, así como la posibilidad de ejecutar consultas federadas en un conjunto preestablecido de fuentes<sup>10</sup>.

El lenguaje de consulta SPARQL describe las primitivas y la sintaxis para especificar una consulta. Las consultas SPARQL se basan en patrones (tripletas) que explotan las relaciones que existen en los datos. Estos patrones triples son similares a las tripletas RDF, excepto que las referencias a recursos se pueden sustituir por variables. Un motor de consultas SPARQL devolvería los recursos para todas las tripletas de la base que cumplen con los patrones definidos en la consulta. Estos patrones pueden ser obligatorios u opcionales, y se pueden establecer conjunciones y disyunciones. En el Ejemplo 2.2 se puede ver una consulta que devuelve el nombre y opcionalmente el correo electrónico de todos los recursos de tipo *foaf:Person*. En el Apéndice B se pueden ver ejemplos de consultas más complejas utilizadas en el contexto de este trabajo.

Ejemplo 2.2: Consulta SPARQL simple.

```
1 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
2 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
3
4 SELECT ?nombre ?email
5 WHERE
6 {
7   ?person rdf:type foaf:Person .
8   ?person foaf:name ?nombre .
9   OPTIONAL {
10    ?person foaf:mbox ?email
11  }
12 }
```

El lenguaje cuenta con cuatro formas básicas de consulta:

- **SELECT**: Se utiliza para extraer resultados en bruto en forma tabular, cada variable seleccionada será devuelta en una columna de la tabla.

<sup>9</sup>Wikidata Query Service: <https://query.wikidata.org/>.

<sup>10</sup>Lista de terminales SPARQL accesibles desde Wikidata: [https://www.mediawiki.org/wiki/Wikidata\\_query\\_service/User\\_Manual/SPARQL\\_Federation\\_endpoints](https://www.mediawiki.org/wiki/Wikidata_query_service/User_Manual/SPARQL_Federation_endpoints)

- **CONSTRUCT**: Se utiliza para devolver la información en un formato de grafo RDF válido.
- **ASK**: Se usa para retornar un resultado de Verdadero/Falso simple. Para esto se verifica si los patrones de la consulta son válidos en los datos.
- **DESCRIBE**: Retorna un grafo RDF a partir de un identificador determinado. Esta pensado para retornar información útil (a criterio del motor de consulta) sobre un recurso específico.

En 2011 se aprobó una nueva versión del lenguaje (SPARQL 1.1) que entre otras cosas agrega soporte para definir sub-consultas y estandariza las operaciones de actualización de datos [14]. Estas últimas se dividen en dos grandes grupos:

- Operaciones de actualización de grafos: básicamente agrega o elimina tripletas de determinados grafos mediante las operaciones: **INSERT**, **DELETE**, **LOAD** y **CLEAR**.
- Operaciones de gestión de grafos: apuntan a la creación y modificación de grafos completos mediante las operaciones: **CREATE**, **DROP**, **COPY**, **MOVE** y **ADD**.

Un aspecto interesante es que SPARQL permite expresar consultas a través de distintas fuentes de datos mediante la cláusula **SERVICE**. Esto da la posibilidad de realizar consultas federadas que combinen datos de diferentes fuentes, siempre y cuando el motor de consultas lo soporte.

Hoy en día existe una gran cantidad de herramientas que implementan las distintas especificaciones SPARQL<sup>11</sup>. Algunos ejemplos son *OpenLink Virtuoso*<sup>12</sup>, *Fuseki*<sup>13</sup>, entre otros. Estas son algunas de las herramientas que hoy en día dan soporte a una gran cantidad de fuentes de datos en la Web Semántica.

## 2.2. Datos Enlazados

Si bien la visión original de la Web Semántica obtuvo un gran desarrollo desde el punto de vista conceptual y teórico, no fueron muchas las aplicaciones de impacto

<sup>11</sup>Lista de implementaciones de SPARQL mantenida por W3C: <https://www.w3.org/wiki/SparqlImplementations>

<sup>12</sup><https://virtuoso.openlinksw.com/>

<sup>13</sup><https://jena.apache.org/documentation/fuseki2/>

desarrolladas. Esto en parte por la poca cantidad de datos disponibles bajo este paradigma. Esto motivó al desarrollo de iniciativas concretas que promuevan la publicación y disponibilidad de datos en la web. De aquí surge el concepto de Datos Enlazados (*Linked Data*).

Los Datos Enlazados se pueden ver como una forma de Web Semántica que se compone de un conjunto de tecnologías y buenas prácticas para publicar y vincular datos estructurados en la web. Es un enfoque más pragmático de la Web Semántica que hace uso de las tecnologías más consolidadas. Apunta a facilitar el acceso y entendimiento de datos que anteriormente se encontraban aislados en las fuentes. En más de una ocasión Tim Berners-Lee ha llamado a los Datos Enlazados como “*la web semántica hecha de la forma correcta*”<sup>14</sup>

En paralelo a la evolución de los Datos Enlazados surge el movimiento de Datos Abiertos. El concepto de Datos Abiertos refiere a los datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona sin restricción alguna más que la de citar su fuente. Un requisito básico para que un dato sea considerado abierto es que se encuentre en un formato legible por máquinas, y esté accesible en Internet bajo una licencia libre. Ambas iniciativas se complementaron a la perfección y dieron lugar al concepto de Datos Abiertos Enlazados (Linked Open Data).

Esta forma de publicación de datos da lugar a un nuevo tipo de web llamada Web de Datos[15]. La web de datos plantea un espacio de datos a nivel global, donde las fuentes de datos son independientes y están distribuidas. La integración se logra gracias a que utilizan un mismo modelo de representación de datos que permite además generar vínculos entre ellas. Esto da la posibilidad de descubrir nueva información siguiendo los enlaces a otras fuentes de datos no contempladas inicialmente.

### 2.2.1. Principios básicos de Datos Enlazados

Como se mencionó anteriormente la idea de los Datos Enlazados es hacer uso de la web para compartir datos estructurados a escala global. Para ello se deben cumplir ciertos principios que aseguran niveles básicos de interoperabilidad. Es por eso que en 2006 se presentan los cuatro principios básicos de los Datos Enlazados<sup>15</sup>:

---

<sup>14</sup><http://www.w3.org/2008/Talks/0617-lod-tbl>

<sup>15</sup><https://www.w3.org/DesignIssues/LinkedData.html>

1. **Usar URIs para identificar cosas.** Se refiere a hacer uso de URIs para identificar de forma única no solo documentos web y contenido digital, sino que además a objetos del mundo real y conceptos abstractos que puedan definirse.
2. **Usar HTTP URIs para que puedan ser dereferenciables.** Este principio restringe a que las URIs sean dereferenciadas bajo el protocolo HTTP con el fin de poder obtener la descripción del objeto o concepto que identifica.
3. **Cuando se accede a dicha URI se debe proporcionar información útil, usando estándares como RDF y SPARQL.** Se debe utilizar el estándar RDF como modelo de datos común para representar los datos. De esta manera cuando se recupere la descripción del recurso previamente dereferenciados se tenga la posibilidad de entenderlo.
4. **Incluir vínculos a otros datos.** Recomienda agregar enlaces a recursos en otros conjuntos de datos con el fin de poder navegar a través de esos enlaces y descubrir nueva información relacionada de forma automática.

Utilizando los principios de Datos Enlazados, es posible consultar y descubrir datos de múltiples fuentes distribuidas y combinarlos sin la necesidad de un esquema común único.

Los Datos Enlazados utilizan dos tecnologías que son parte fundamentales de la web actual: **Uniform Resource Identifiers (URI)** y **HyperText Transfer Protocol (HTTP)**.

Las URIs proporcionan una forma simple de crear nombres únicos a escala mundial de manera descentralizada. De esta forma cualquier organización o persona puede escoger el esquema de URIs que identificarán sus datos. A diferencia de la web actual donde las URIs hacen referencia a documentos web o contenidos multimedia, las URIs en los Datos Enlazados representan objetos del mundo real o conceptos abstractos. Estas entidades (denominadas recursos) se identifican por URIs que utilizan el esquema `http://` lo que posibilita que pueden ser buscadas y accedidas mediante protocolo HTTP, por eso son llamadas HTTP URIs. [15]. La acción de buscar y recuperar la descripción de un recurso a partir de su URI se denomina dereferenciación. Normalmente este proceso implica una negociación entre las partes (el consumidor y el proveedor de los datos) donde se acuerda el formato en el que se espera recibir la representación del recurso.

Así como la web de documentos permite enlazar documentos mediante hipervínculos, con RDF se pueden enlazar recursos de diferentes conjuntos de datos. Para

esto basta con crear una tripleta cuyo objeto sea un recurso en el espacio de nombres de otro conjunto de datos. En el Ejemplo 2.1 se puede ver un caso de este tipo en la propiedad *dbo:birthPlace* del sujeto *Luca*, ya que apunta a un recurso que representa a la ciudad de Montevideo<sup>16</sup> que se encuentra en la fuente de datos DBpedia. Siguiendo este enlace se podría obtener más información sobre este recurso almacenada en DBpedia, como su uso horario, idioma, o incluso una lista de personas destacadas que nacieron en dicho lugar. De esta manera se va conformando una web de Datos Enlazados de fuentes distribuidas e independientes.

### 2.2.2. Vocabularios para describir los datos

Los vocabularios sirven para clasificar cosas, establecer relaciones o definir restricciones útiles para describir un determinado dominio de conocimiento. También están descritos en RDF y hacen uso de términos de otras especificaciones tales como **RDFS** y **OWL**, que sientan las bases para su construcción. Los vocabularios se pueden ver como una ontología que contiene únicamente los elementos del TBox. Sin embargo en el contexto de la Web Semántica muchas veces suelen confundirse ambos conceptos y usarse de forma indistinta<sup>17</sup>.

Se han desarrollado una gran cantidad de vocabularios para distintos propósitos, y dado que se basan en las mismas tecnologías y principios de los Datos Enlazados, cualquier persona u organización puede crear sus propios vocabularios y dejarlos disponibles en la web para que sean usados. En este sentido la recomendación es tratar de reutilizar los ya existentes siempre que sea posible, esto facilita la utilización e integración de datos de múltiples fuentes[15]. Existen varias iniciativas interesantes que facilitan la búsqueda y el descubrimiento de nuevos vocabularios, como por ejemplo *LOV.org*<sup>18</sup>, *Prefix.cc*<sup>19</sup> o *Schema.org*<sup>20</sup>. A continuación se listan algunos de los vocabularios más populares relacionados con el área de investigación de esta tesis:

- Dublin Core Metadata Initiative (DCMI). Es uno de los vocabularios más usados y define atributos o metadatos generales tales como *título*, *creador*, *fecha* o *asunto*.
- Friend-of-a-Friend (FOAF): es un vocabulario que define términos para des-

---

<sup>16</sup><http://dbpedia.org/page/Montevideo>

<sup>17</sup><https://www.w3.org/standards/semanticweb/ontology>

<sup>18</sup><http://lov.okfn.org/dataset/lov/>

<sup>19</sup><http://prefix.cc>

<sup>20</sup><http://schema.org>



cribir personas y organizaciones, sus actividades y la relación que tienen con otras personas y objetos.

- Creative Commons (CC): define términos para describir licencias de copyright en RDF.
- Semantically-Interlinked Online Communities (SIOC): está diseñado para describir aspectos relacionados con comunidades en línea, tales como usuarios, foros, etc.
- Bibliographic Ontology (BIBO): provee conceptos y propiedades para describir referencias bibliográficas.
- OAI Object Reuse and Exchange: es usado por varias bibliotecas y sistemas de publicaciones electrónicas para representar agregación de recursos tales como las diferentes ediciones de documentos o su estructura interna.
- Review Vocabulary: describe críticas y ratings de productos o servicios.
- Basic Geo (WGS84): provee términos tales como latitud y longitud para describir ubicaciones geográficas.
- SKOS: es un modelo común para compartir y enlazar sistemas de organización de conocimiento en la web. La idea se basa en compartir estructuras similares y permite representar tesauros, taxonomías y jerarquías temáticas en general[16].

Si bien un vocabulario requiere del trabajo conjunto y el consenso de comunidades que conozcan el dominio objetivo, no deja de ser una visión parcial de la realidad creada por un conjunto de personas. Esto plantea un gran desafío a la hora de integrar datos de distintas fuentes y dominios. Por tal motivo existen varias iniciativas que trabajan en la línea de la armonización de metadatos y vocabularios, que buscan facilitar dicha integración[17].

### **2.2.3. Ciclo de vida**

La publicación de Datos Enlazados puede verse como un proceso de gestión de información que va desde su obtención hasta su uso final. Este proceso consta de varias actividades que requieren la toma de múltiples decisiones, muchas de ellas relacionadas con aspectos tecnológicos y de diseño de la solución. En la literatura se pueden encontrar varias propuestas metodológicas que abordan este

proceso. Muchas de ellas están basadas en experiencias empíricas de proyectos de publicación de datos.

A continuación se listan las etapas de tres metodologías mencionadas en el reporte de buenas prácticas para la publicación de Datos Enlazados de la W3C[18]:

- (1) Identify, (2) Model, (3) Name, (4) Describe, (5) Convert, (6) Publish, (7) Maintain. [19]
- (1) Data awareness, (2) Modeling, (3) Publishing, (4) Discovery, (5) Integration, (6) Use cases.[20]
- (1) Specify, (2) Model, (3) Generate, (4) Publish, (5) Exploit. [21]

En términos generales se puede ver que, aunque varíe su nombre o alcance, todas las propuestas comparten muchas de las actividades y objetivos. En algunos casos las tareas se dividen y en otros se combinan o solapan. A continuación se describen a nivel general algunas de estas etapas:

**Identificación/Especificación.** El primer paso es identificar y analizar las fuentes de datos. La idea aquí es identificar con que datos se cuenta, cual es la naturaleza de los mismos y quienes son los principales involucrados. En proyectos vinculados a datos de gobierno[21] esta etapa suele usarse también para establecer acuerdos entre partes así como para planificar y estimar costos del proyecto.

**Diseño de URIs.** En esta etapa se define el esquema de URIs, es decir la forma de nombrar a cada recurso dentro del conjunto de datos. El esquema definido debe ser simple desde el punto de vista nemotécnico, estable a lo largo del tiempo y manejable desde el punto de vista del volumen y evolución de los datos[22].

**Modelado.** En esta etapa se analizan los datos desde el punto de vista semántico. Esto implica seleccionar cuáles son los vocabularios y ontologías que mejor se ajustan al significado de los datos. Se busca en lo posible reutilizar vocabularios existentes, aunque podría requerirse el diseño de nuevos.

**Generación/Conversión.** En esta etapa se transforman los datos seleccionados a grafos RDF. Esta etapa puede requerir la limpieza de los datos o la selección de un formato de serialización RDF determinado. Otra actividad importante de esta etapa es la detección y creación de vínculos a otros juegos de datos. Esto puede implicar el reconocimiento de entidades en los datos y la resolución de identidad.

**Publicación.** En esta etapa se deja disponible en la web el grafo RDF previamente

generado. Esto implica tener el mecanismo de dereferenciación de las URIs activo para retornar información relacionada con el recurso solicitado. Para facilitar la reutilización de los datos es recomendable proporcionar una terminal SPARQL y agregar metadatos que describan el conjunto de datos completo. En algunos casos suele publicarse también un archivo comprimido con todos los datos para descargar, denominado *dump*.

**Explotación/Usó.** Esta etapa refiere al uso de los datos publicados y puede ser tan diverso como las posibles ideas que puedan existir. Puede ir desde el desarrollo de herramientas hasta el análisis de datos. Algunas de estas aplicaciones pueden ser motores de búsqueda semánticos, navegadores de Datos Enlazados o aplicaciones específicas de dominio. El uso puede implicar la integración con otras fuentes (ej. *mashups*) y en caso de volver a publicarse podrían enriquecer aún más los datos originales en un proceso de mejora continua. En [19] se distingue una actividad especial dedicada al mantenimiento y se resalta la importancia de contar con datos actualizados que favorezcan su uso por terceros.

Otra visión interesante del ciclo de vida es la planteada en el proyecto LOD2 Stack[23]. LOD2 es un proyecto de integración de datos a gran escala co-financiado por la Comisión Europea dentro del Programa de Trabajo sobre Tecnologías de la Información y la Comunicación. Se centra en identificar las distintas tareas involucradas en la manipulación de los datos a lo largo de la vida de los mismos (figura 2.4). El proyecto va más allá de la definición de actividades y propone un stack completo de tecnologías y herramientas para dar soporte a cada una de ellas.

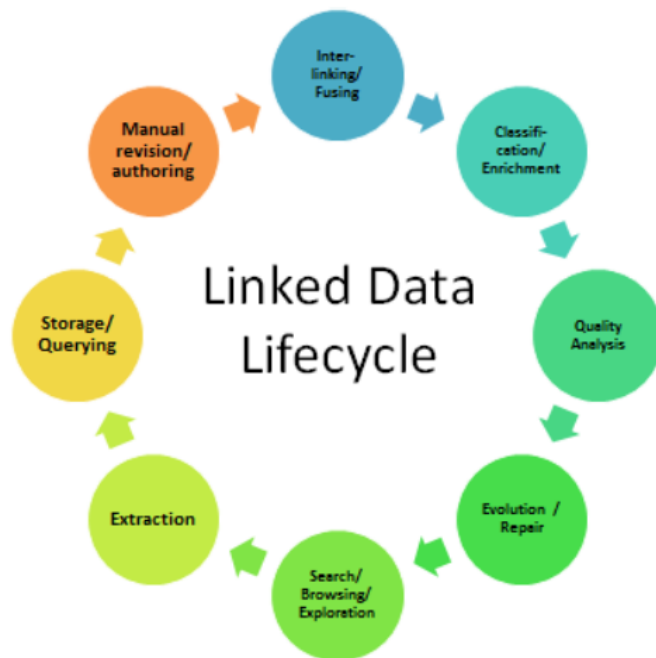


Figura 2.4: Ciclo de vida planteado en LOD2 Stack

Es importante mencionar que no todas las etapas aplican a todos los escenarios y no siempre el punto de partida es el mismo. De todas formas, y al igual que otros procesos de análisis y gestión de información, la publicación de Datos Enlazados requiere un esfuerzo considerable. Por tal motivo es natural que el equipo de trabajo ponga mayor énfasis en los aspectos relevantes a su área de conocimiento o a sus requerimientos particulares. Por ejemplo destinando mayor esfuerzo en describir con mayor especificidad un determinado subconjunto de los datos, o enlazando los datos únicamente con ciertas fuentes de interés. Este hecho reafirma la visión de un proceso incremental, ya que el ciclo de vida de los Datos Enlazados se puede ver como una iteración sobre las distintas etapas, con el fin de ir enriqueciendo paulatinamente la semántica del conjunto de datos. El enfoque seleccionado va a depender fundamentalmente de la naturaleza de los datos, la disponibilidad y madurez de las herramientas y la experiencia del equipo de trabajo.

## 2.2.4. Buenas prácticas

Además de las metodologías planteadas en la sección anterior existen varias iniciativas que recopilan patrones y buenas prácticas que sirven de insumo a la hora de tomar decisiones en el proceso de publicación.

Un ejemplo destacado es la guía de buenas prácticas[18] creada por el grupo de trabajo de Datos Abiertos Enlazados de Gobierno de la W3C. A continuación se comentan las mejores prácticas mencionadas en dicha guía:

1. **Preparar a los involucrados.** Se debe explicar cual es el concepto de Datos Enlazados, y en qué consiste el proceso de publicación y mantenimiento.
2. **Seleccionar el conjunto de datos.** Se debe tratar de elegir los juegos de datos que tengan mayor valor, teniendo en cuenta por ejemplo la popularidad o el potencial de re-uso.
3. **Modelar los datos.** Implica representar los conceptos identificados en los datos de forma independiente de la aplicación. En este caso se recomienda que sea un equipo multidisciplinario que cuente con personas que conozcan del dominio en cuestión, para que describan las entidades y relaciones implícitas en los datos.
4. **Especificar una licencia apropiada.** En lo posible utilizar una licencia abierta. El contar con información de propiedad y términos de uso claros favorece el re-uso. Un recurso valioso en este sentido son la familias de licencias de Creative Commons<sup>21</sup>.
5. **Definir un buen esquema de URIs.** A la hora de definir las URIs se debe tener en consideración, el soporte multilinguaje, la evolución de los datos a lo largo del tiempo y la estrategia de persistencia de los datos.
6. **Utilizar vocabularios estándares.** Se debe reutilizar los vocabularios existentes siempre que sea posible. Se pueden extender vocabularios estándares si es necesario. La creación de nuevos vocabularios debe aplicarse únicamente cuando sea estrictamente necesario y siguiendo las buenas prácticas.
7. **Convertir los datos.** Convertir los datos a una representación determinada. Esto normalmente se realiza mediante procesos automatizados. Se recomienda incluir metadatos que describan el conjunto de datos y enlaces a otros

---

<sup>21</sup><http://creativecommons.org/>

juegos de datos.

8. **Proveer acceso de máquina a los datos.** Proveer varias formas de acceso para que procesos automatizados accedan a los datos mediante los mecanismos web estándar. Por ejemplo, resolución directa de URIs, terminal SPARQL, o descarga de un archivo (o *dump*) con el conjunto completo de datos.
9. **Anunciar los nuevos conjuntos de datos.** Difundir los nuevos conjuntos de datos en los ámbitos adecuados utilizando varios canales de comunicación, por ejemplo, listas de correo, blogs o boletines de noticias.
10. **Reconocer el contrato social.** Una vez que se publiquen, se debe reconocer la responsabilidad de mantenimiento de los datos. Se debe asegurar la disponibilidad a lo largo del tiempo. Esto es clave para generar confianza en los usuarios y asegurar la continuidad de posibles aplicaciones que hacen uso de los datos.

### 2.2.5. Patrones de publicación

Respetar los principios de los Datos Enlazados no implica dejar de utilizar los sistemas de información o aplicaciones de negocio existentes, sino agregar una capa más que se encargue de implementar un determinado proceso de transformación y publicación de los datos. Uno de los principales aspectos a tener en cuenta a la hora de definir dicho proceso es la naturaleza de los mismos. En [24] se plantean tres posibles puntos de partida de distinta naturaleza:

- Partiendo de datos estructurados con algún mecanismo de consulta. Por ejemplo una base de datos relacional.
- Partiendo de datos estructurados sin mecanismo de consulta. Por ejemplo documentos XML o CSV.
- Partiendo de documentos de texto libre, sin estructura.

Como se muestra en la Figura 2.5 cada uno de estos caminos puede requerir distintas estrategias y herramientas para la preparación, almacenamiento y publicación de los datos. Otros aspectos que pueden incidir en la solución adoptada son el volumen y la frecuencia de actualización de los datos. Por ejemplo, si el volumen de datos a publicar es reducido lo más conveniente podría ser la publicación de un

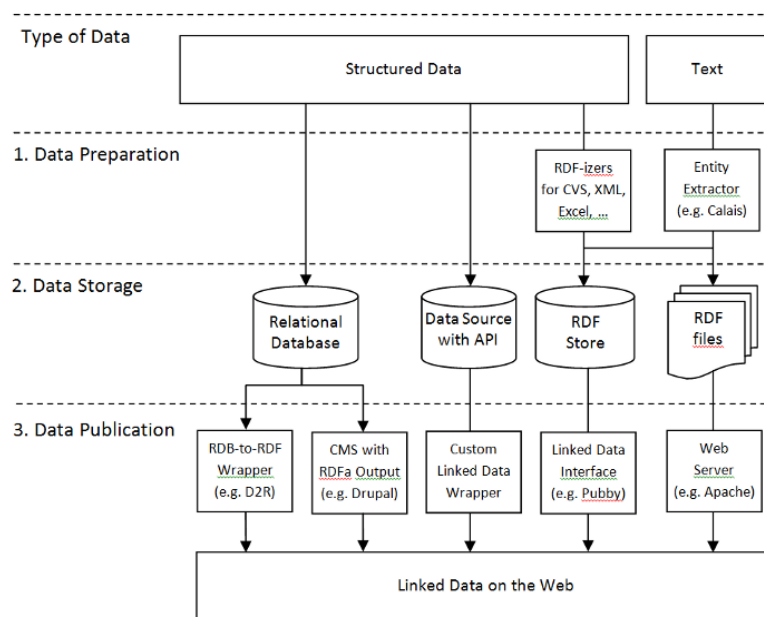


Figura 2.5: Patrones de publicación según la naturaleza de los datos [24]

único documento RDF estático. Por el contrario, si se debe publicar un conjunto de datos voluminoso con muchas entidades, lo recomendable sería separarlo en un archivo por entidad. Lo mismo ocurre con juegos de datos que cambian frecuentemente, en estos casos seguramente se requiera de soluciones más complejas, que faciliten la gestión y generación de los datos de forma eficiente.

Otra iniciativa destacada es el catálogo de patrones para el modelado, publicación y consumo de Datos Enlazados[25]. El catálogo abarca una variedad amplia de actividades diferentes, desde el diseño de identificadores al desarrollo de aplicaciones que hacen uso de los datos. El objetivo de los autores es crear una lista de referencia que sea útil tanto para el principiante como para las personas experimentadas. Los patrones se organizan en cinco categorías diferentes: *identificación*, *modelado*, *publicación*, *administración de datos*, y *aplicación*. Cada patrón está compuesto por un breve contexto que explica el problema y las posibles soluciones. También presenta ejemplos, discusiones y recursos relacionados que enriquecen aun más cada caso.

## 2.2.6. Estado actual de la Web de Datos

La iniciativa de Datos Abiertos Enlazados logró promover fuertemente la publicación de datos, lo que redundó en un aumento sostenido en la publicación de Datos Enlazados en la última década. Esto motivó el surgimiento de varias iniciativas que realizan análisis estadísticos de la web de datos, y nos ayudan a comprender mejor su evolución en términos del volumen, la cobertura y la calidad de los datos publicados. En este sentido proyectos como *LODStats*, *SPARQLES* o *LODCloud* sirven de referencia a la hora de analizar el estado actual de la web de datos.

*LODStats*[26] es un proyecto que monitorea de forma continua cerca de 10000 conjuntos de datos disponibles en diversos catálogos de datos, como por ejemplo Data Hub<sup>22</sup>. Pone a disposición en su sitio web<sup>23</sup> y en forma de Datos Enlazados, información estadística detallada sobre las clases, propiedades, links, idiomas y los vocabularios utilizados que componen cada juego de datos.

Por otro lado *SPARQLES* (SPARQL Endpoint Status) es una iniciativa de supervisión de terminales SPARQL. *SPARQLES* supervisa la disponibilidad, el rendimiento y la interoperabilidad de los distintos puntos de entrada. El proyecto proporciona un sitio web<sup>24</sup> y una API de consulta para acceder a dichos datos[27].

Finalmente *LOD cloud* es una iniciativa muy conocida gracias al diagrama de nube de Datos Abiertos Enlazados. En la figura 2.6 se puede ver la última versión de este diagrama. El diagrama se publica periódicamente en el contexto del proyecto y muestra cada juego de datos como un círculo. El color del círculo representa el área temática mientras que el tamaño representa la cantidad de vínculos desde otros juegos de datos. Esto se puede ver como una medida de relevancia del juego de datos. A diferencia de *LODStat*, además de analizar catálogos de datos, su estudio recolecta datos mediante un sistema de *crawler*, lo que proporciona un universo de estudio potencialmente mayor.

---

<sup>22</sup><http://datahub.io/>

<sup>23</sup><http://lodstats.aksw.org>.

<sup>24</sup><http://sparqles.ai.wu.ac.at>



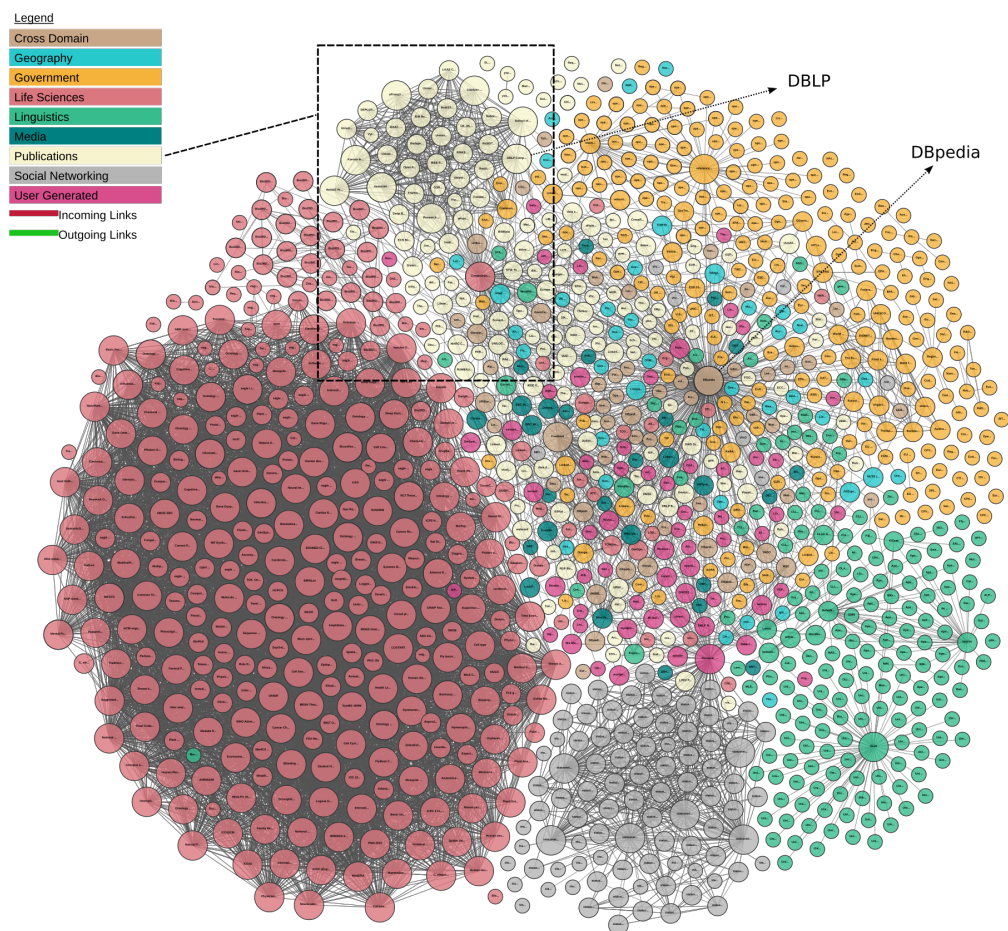


Figura 2.6: Grafo con la estructura general y la categorización por dominio temático de los juegos de datos publicados. El tamaño de lo círculos representa el grado de referencias que recibe desde otros juegos de datos.<sup>25</sup>

El último análisis detallado se publicó en abril de 2014 [28]. En el dominio de Publicaciones los juegos de datos más referenciados son la lista de encabezamiento de materias (LEM) de la Biblioteca del Congreso de Estados Unidos (LCSH), y de la biblioteca de Alemania (DNB). En este dominio los tres vocabularios más usados (excluyendo rdf, rdfs and owl) son: dct (83,91%), foaf (69,23%) y bibo (41,67%). A su vez el 34,38% de los juegos usa vocabularios propietarios, de los cuales el 69,49% no son dereferenciables. En la tabla 2.1 se muestra un resumen

<sup>25</sup>Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. <http://lod-cloud.net/>

del cumplimiento de algunas de las prácticas recomendadas a la hora de publicar datos.

Tabla 2.1: Resumen cumplimiento de buenas prácticas generado a partir de los datos publicadsos en [28]

Recomendación	% total	% dom. publicación	notas
Agregar información de procedencia	36,69 %	40,63 %	94,87 % usa dublin core
Proveer información de licencia	9,96 %	4,17 %	dc/dct:license (7,39 %), cc:license (2,07 %) and dc/dct:rights (1,68 %)
Proveer información sobre el dataset	14,69 %	13,54 %	
Proveer formas alternativa de acceso	11,14 %	13,54 %	SPARL (12.50 %), Dump (4,17 %)

Algunas de las conclusiones del segundo estudio [28] indican que el número de juegos de datos se duplicó entre 2011 y 2014. También se notó un incremento en la adopción de vocabularios conocidos y estándares (tales como FOAF), y un descenso de dos tercios en el uso de vocabularios propietarios. Un aspecto negativo que se mantuvo incambiado es que la información de procedencia y licenciamiento raramente es proporcionada por las fuentes de datos. Esto último plantea dificultades a la hora de reutilizar esa información en publicaciones electrónicas, ya sea por el riesgo de infringir leyes de copyright o simplemente por el desconocimiento de la fuente que proviene.

# Capítulo 3

## Aplicación

Con el fin de explorar estos temas se plantearon dos escenarios de publicación de datos. Ambos casos se enmarcan en el dominio de las publicaciones de índole académico, por tal razón manejan conceptos similares. Por otro lado presentan características diferentes desde el punto de vista de la naturaleza de los datos y su forma de acceso. Un aspecto fundamental es que en ambos casos se dispone de libre acceso a los datos.

El primer caso corresponde al análisis de la plataforma de creación y publicación de libros de texto CNX.org. Su elección fue pautaada por ser una de las plataformas candidatas para ser usada en el proyecto LATin<sup>26</sup>. Dispone de una gran cantidad de libros accesibles por medio de distintos mecanismos y formatos.

El segundo caso corresponde al análisis de la producción bibliográfica de los docentes del **Instituto de Computación** de la **Facultad de Ingeniería (UdelaR)**. A diferencia del caso anterior no se cuenta con una única fuente de datos, por lo que se hace necesario la integración de diferentes fuentes para poder obtener un conjunto lo más completo posible de publicaciones.

Los casos se presentan siguiendo las etapas del ciclo de vida planteada en la sección 2.2.3. Para facilitar la lectura se usaron prefijos para las URIs más utilizadas. En la Tabla 3.1 se puede ver la lista de prefijos mencionados.

---

<sup>26</sup><http://proeva.udelar.edu.uy/institucional/proyectos/latin/>

Tabla 3.1: Lista de prefijos de URIs utilizados a lo largo del documento

Prefijo	URI
bibtex	<a href="http://purl.org/net/nknouf/ns/bibtex#">http://purl.org/net/nknouf/ns/bibtex#</a>
dblp	<a href="http://dblp.l3s.de/d2r/resource/authors/">http://dblp.l3s.de/d2r/resource/authors/</a>
dbc	<a href="http://dbpedia.org/resource/Category:">http://dbpedia.org/resource/Category:</a>
dbo	<a href="http://dbpedia.org/ontology/">http://dbpedia.org/ontology/</a>
dbr	<a href="http://dbpedia.org/resource/">http://dbpedia.org/resource/</a>
dct	<a href="http://purl.org/dc/terms/">http://purl.org/dc/terms/</a>
dc	<a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a>
doc	<a href="http://www.w3.org/2000/10/swap/pim/doc#">http://www.w3.org/2000/10/swap/pim/doc#</a>
docente	<a href="http://data.fing.edu.uy/schema/docentes/">http://data.fing.edu.uy/schema/docentes/</a>
foaf	<a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a>
owl	<a href="http://www.w3.org/2002/07/owl#">http://www.w3.org/2002/07/owl#</a>
rdfs	<a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a>
rdf	<a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://www.w3.org/1999/02/22-rdf-syntax-ns#</a>
swrc	<a href="http://swrc.ontoware.org/ontology#">http://swrc.ontoware.org/ontology#</a>
vcard	<a href="http://www.w3.org/2001/vcard-rdf/3.0#">http://www.w3.org/2001/vcard-rdf/3.0#</a>
xsd	<a href="http://www.w3.org/2001/XMLSchema#">http://www.w3.org/2001/XMLSchema#</a>

### 3.1. CNX.org

CNX<sup>27</sup> es una organización sin fines de lucro formada por un consorcio de universidades. Está liderada por la Universidad Rice<sup>28</sup> y su objetivo principal es proveer soporte a la creación y divulgación de libros de texto digitales gratuitos. Para la creación y gestión de los materiales educativos CNX utiliza una plataforma de código abierto denominada Rhaptos<sup>29</sup>. Una de las características principales de Rhaptos es su capacidad para la creación de libros de texto de forma colaborativa. Provee funcionalidades que facilitan la reutilización y combinación de contenidos existentes en la plataforma y la revisión por pares. En este contexto se busca dejar disponible como Datos Enlazados los metadatos que describen a los contenidos de la plataforma.

<sup>27</sup>[http://cnx.org/aboutus/index\\_html](http://cnx.org/aboutus/index_html)

<sup>28</sup><http://www.rice.edu/>

<sup>29</sup><https://github.com/Rhaptos/>

### 3.1.1. Especificación

En esta etapa se analiza la fuente de datos y las posibilidades de acceso a los mismos que brinda la plataforma, así como los metadatos disponibles en cada caso.

La unidad mínima de contenido que maneja CNX se denomina módulo que son posteriormente agregados en colecciones. Las colecciones pueden ser libros de texto, cursos, revistas o cualquier otro tipo de publicación digital que se desee representar. Todo el material generado en CNX es licenciado bajo [Creative Commons Attribute License \(CC-BY\)](#) lo que favorece su reutilización. A lo largo del documento nos vamos a referir a las colecciones y módulos con el término *contenido*.

Los módulos y colecciones pueden ser creados directamente en la plataforma utilizando un editor rico, o importados mediante archivos en otros formatos, como por ejemplo *OpenOffice* o  $\text{\LaTeX}$ . Una vez creados son almacenados de forma nativa en XML utilizando los formatos [CNXML](#) y [CollXML](#) respectivamente. Ambos formatos tienen una sección de metadatos especificada con el formato [MDML](#)<sup>30</sup>. El ejemplo 3.1 muestra un fragmento de la sección de metadatos de un módulo particular.

Ejemplo 3.1: fragmento de la sección de metadatos en formato MDML para un módulo determinado

```
11 <md:title>The Sampling Theorem</md:title>
12 <md:version>2.20</md:version>
13 <md:created>2000/07/27</md:created>
14 <md:revised>2012/07/06 18:20:37.283 GMP-5</md:revised>
15 <md:actors>
16   <md:person userid="carolrb">
17     <md:firstname>Carol</md:firstname>
18     <md:surname>Bettoney</md:surname>
19     <md:fullname>Carol Bettoney</md:fullname>
20     <md:email>carolrb@alumni.rice.edu</md:email>
21   </md:person>
22   <md:person userid="dhj">
23     <md:firstname>Don</md:firstname>
24     <md:surname>Johnson</md:surname>
25     <md:fullname>Don Johnson</md:fullname>
26     <md:email>dhj@rice.edu</md:email>
27   </md:person>
28 </md:actors>
29 <md:roles>
30   <md:role type="author">dhj</md:role>
31   <md:role type="maintainer">dhj carolrb</md:role>
32   <md:role type="licensor">dhj</md:role>
33 </md:roles>
```

<sup>30</sup><http://legacy.cnx.org/help/authoring/xml>

## Metadatos

Los módulos y colecciones de CNX soportan un conjunto básico de metadatos especificado con el formato MDML. En la Tabla 3.2 se puede observar la lista completa de metadatos disponibles.

Se evaluó la posibilidad de extensión del conjunto de metadatos que proporciona la plataforma y se realizaron las siguientes observaciones:

1. Los metadatos son fijos y se validan contra el esquema MDML cada vez que se guarda la colección o el módulo.
2. El formulario web de carga de metadatos es fijo, no hay forma de agregar nuevos campos o propiedades<sup>31</sup>
3. Si bien la plataforma provee una manera de editar directamente el código fuente de las colecciones y módulos (CNXML y CollXML), en la documentación se especifica que la sección de metadatos es de solo lectura<sup>32</sup> y debe editarse desde el formulario de carga mencionado en el segundo punto.

Por lo dicho anteriormente soportar nuevos metadatos que describan a los módulos y colecciones requeriría (1) cambiar el esquema MDML que los especifica (2) cambiar el comportamiento de la plataforma para permitirle al usuario modificar el código fuente de la sección de metadatos, permitiendo así complementar los metadatos con otros esquemas.

## Mecanismos de acceso

La plataforma provee varias formas de acceso a los datos<sup>33</sup>. Algunas orientadas a usuarios finales, tales como la navegación en línea de los contenidos en el sitio web, o la descarga de formatos tales como PDF o EPUB. A los efectos de nuestros intereses es necesario poder acceder a los metadatos de forma sistemática con el

---

<sup>31</sup><http://cnx.org/content/m19610/latest/module-metadata.png>

<sup>32</sup><http://legacy.cnx.org/help/authoring/xml#mdml>

<sup>33</sup>[https://github.com/openstax/openstax\\_api](https://github.com/openstax/openstax_api)

Tabla 3.2: Metadatos soportados en el formato MDML

Propiedad	Obligatorio	Ejemplo
repository	S	http://cnx.org/content
content-url	S	http://cnx.org/content/col10373/1.2
contentId	S	col10373
title	S	Señales y Sistemas
short-title		
subtitle		
version	S	1.2
created	S	"2006/08/22 17:16:52.856 GMT-5"
revised	S	"2006/09/28 12:13:22.030 GMT-5"
license	S	"http://creativecommons.org/licenses/by/2.0/"
derived-from		"http://cnx.org/content/col10064/1.6"
subjectlist		["Science and Technology"]
keywords		["Displacement", "Distance", "Kinematics", "Motion", "Position"]
abstract		Este curso trata acerca de señales, sistemas, y transformadas a partir de las bases matemáticas y teóricas hasta las implementaciones prácticas en circuitos y algoritmos. Al terminar ELEC 301, tendrá un amplio entendimiento de las matemáticas y temas prácticos relacionados con las señales en tiempo continuo y tiempo discreto, sistemas lineales invariantes en el tiempo, la convolución, y la transformada de Fourier.
language	S	es
institution <sup>a</sup>		Rice University
course-code <sup>a</sup>		Elec 301
instructor <sup>a</sup>		Richard Baraniuk
homepage <sup>a</sup>		
extended-attribution		
education-levellist		
objectives		
col-homepage <sup>a</sup>		
Roles disponibles <sup>b</sup>		
author		[richb*]
maintainer		["pedros", richb*, cmpotes*]
licensor		[richb*]
Metadatos asociados a un rol de tipo persona		
id	S	richb
firstname	S	Ricardo
othername		Anthony
surname	S	Radaelli-Sanchez
fullname	S	Ricardo Radaelli-Sanchez
email		ricky@alumni.rice.edu
homepage		alumni.rice.edu/ ricky
Metadatos asociados a un rol de tipo organización		
id	S	riceu
fullname	S	Rice University
shortname	S	Rice University
email		contact@rice.edu
homepage		http://www.rice.edu

<sup>a</sup>Usado únicamente en colecciones

<sup>b</sup>Los roles pueden hacer referencia a Personas u Organizaciones

objetivo de facilitar el procesamiento necesario para la generación automática de Datos Enlazados. En este contexto se destacan tres mecanismos de acceso provistos por la plataforma que se describen a continuación.

**Descarga de los contenidos completos en su formato nativo (CNXML y CollXML).** Cada contenido presenta un enlace para descargar el código fuente en su formato nativo. Esto nos da la posibilidad de acceder al conjunto completo de metadatos. En contrapartida se debe buscar algún otro mecanismo para acceder a la lista completa de los contenidos almacenados en la plataforma. Una posible solución podría incluir la implementación de un **crawler** que recorra periódicamente el sitio web descargando cada contenido que encuentre para su posterior procesamiento. También se podría utilizar en combinación con otros mecanismos como el protocolo OAI-PMH o el **RSS** para obtener el listado total de contenidos o las futuras actualizaciones<sup>34</sup>.

**Consulta Mediante API Web Service.** Otra posibilidad es utilizar la interfaz de servicios REST provista por la plataforma para acceder a gran parte de los datos y metadatos almacenados. Algunos de esos servicios permiten acceder a las imágenes, archivos audio/video y otros tipos de recursos embebidos en un módulo. También permite acceder a la estructura interna de las colecciones listando las sub-colecciones o módulos que la componen. Desde el punto de vista de los metadatos, provee acceso a un sub-conjunto del total de los metadatos soportados por la plataforma. Una desventaja importante es que la documentación de la **API** esta bastante desactualizada. Algunas operaciones figuran como propuestas para futuras implementaciones y otras simplemente no funcionan. Un ejemplo de esto es la operación *getKeywords* la cual esta especificada en la documentación pero no esta implementada. Por otro lado, existen operaciones implementadas que no están documentadas tal como *getSubjects*. A esto se le suma el mismo inconveniente del formato nativo ya que no se cuenta con una operación para listar todos los contenidos. Todas las operaciones se aplican sobre un contenido determinado para lo cual se debe conocer su identificador de antemano.

**Cosecha Mediante protocolo OAI-PMH.** Por último la plataforma provee acceso a un subconjunto de los metadatos por medio del protocolo OAI-PMH. Este protocolo está orientado a la extracción de metadatos. Funciona sobre el protocolo HTTP y se compone de seis acciones diferentes para la consulta de datos: *Identify*, *ListMetadataFormats*, *ListSets*, *ListIdentifiers*, *ListRecords*, *GetRecord*. CNX soporta de forma nativa tres formatos diferentes de intercambio [29]<sup>35</sup>:

---

<sup>34</sup><http://legacy.cnx.org/content/recent.rss>

<sup>35</sup><http://legacy.cnx.org/content/OAI?verb=ListMetadataFormats>



- [http://www.imsglobal.org/xsd/imsmd\\_v1p2](http://www.imsglobal.org/xsd/imsmd_v1p2)
- [http://www.openarchives.org/OAI/2.0/oai\\_dc/](http://www.openarchives.org/OAI/2.0/oai_dc/)
- [http://cnx.rice.edu/cnx\\_dc/](http://cnx.rice.edu/cnx_dc/)

La utilización de este protocolo presenta varias ventajas:

- Resuelve la interacción con el proveedor del servicio para el listado de nuevos contenidos y el acceso a los metadatos.
- Es un estándar muy usado. Actualmente existen más de 2000 instituciones que proporcionan acceso a sus bases de datos bibliográficas por medio de este protocolo<sup>36</sup>.
- Existe una gran cantidad de herramientas desarrolladas que lo soportan, tanto para proveer el servicio, como para consumir y procesar los datos<sup>37</sup>.
- El protocolo es independiente del formato de intercambio. Esto da la posibilidad de agregar nuevos metadatos en el futuro creando un nuevo formato de intercambio (o extendiendo uno ya existente) sin necesidad de cambiar el protocolo de acceso.

## Conclusión

En la Tabla 3.3 se muestra una lista detallada de la disponibilidad de los distintos metadatos según los diferentes mecanismos de acceso disponibles. Para el caso del protocolo OAI-PMH se discriminó también según el formato de intercambio utilizado. Como se puede observar cada mecanismo permite acceder a distintos sub-conjuntos de metadatos. Si bien no existe una gran diferencia entre dichos sub-conjuntos, ninguno brinda acceso al conjunto total de metadatos, salvo el acceso al formato nativo XML (no incluido en la tabla).

Una posibilidad es acceder directamente a los fuentes de cada contenido mediante la implementación de un **crawler**. Esto nos daría acceso total a los datos y metadatos con la contrapartida de requerir un costo de implementación mayor y menores posibilidades de reutilización con otras fuentes.

Por otro lado el uso de la API REST y el protocolo OAI-PMH cuentan con limi-

<sup>36</sup><http://www.openarchives.org/Register/BrowseSites>

<sup>37</sup><http://www.openarchives.org/pmh/tools/tools.php>

Tabla 3.3: Mecanismos y nivel de acceso a metadatos

Tipo de Acceso →	WEB	WebService REST	OAI-PMH		
Propiedad ↓	Interfaz WEB <sup>a</sup>	Operación	oai_dc	cnx_dc	ims_1_2_1 <sup>b</sup>
repository	R	-	-	-	-
content-url	R	-	-	-	-
content-id	R	-	<identifier>	<identifier>	<identifier> <sup>c</sup>
title	R/W	getTitlecol	dc:title	dc:title	ims1_2_1:title
short-title	-	-	-	-	-
subtitle	-	-	-	-	-
version	R	getVersion	-	-	ims1_2_1:version
created	R	getCreated	-	-	-
revised	R	getRevised	dc:date	dc:date	ims1_2_1:date <sup>d</sup>
license	R	getLicense	dc:rights	dc:rights	ims1_2_1:rights
derived-from	R	-	-	-	-
subjectlist	R/W	-	-	cnxdc:cnx-subject	-
keywords	R/W	getKeywords	dc:subject	dc:subject	ims1_2_1:keyword
abstract	R/W	getAbstract	dc:description	dc:description	ims1_2_1:description
language	R/W	-	dc:language	dc:language	ims1_2_1:language
institution	R/W (col)	getInstitution	-	-	-
course-code	R/W (col)	getCode	-	-	-
instructor	R/W (col)	-	-	-	-
homepage	R/W (col)	getHomepage	-	-	-
extended-attribution	-	-	-	-	-
education-levellist	-	-	-	-	-
objectives	-	-	-	-	-
col-homepage	-	-	-	-	-
author	R/W	getAuthors (id)	dc:creator (fullname)	dc:creator (fullname)	ims1_2_1:contribute (vCard) <sup>e</sup>
maintainer	R/W	getMaintainers (id)	-	cnxdc:maintainer (fullname)	-
licensor	R/W	getLicensors (id)	-	cnxdc:maintainer (fullname)	-

<sup>a</sup>Indica si el metadato se puede leer (R) o escribir (W) desde el sitio web proporcionado por la plataforma. (col) refiere a que solo está accesible en contenidos de tipo Colección.

<sup>b</sup>El atributo utilizado en esta columna puede no ser exacto ya que en muchos casos la estructura XML del formato ims121 utiliza varios elementos anidados para representar dicha propiedad

<sup>c</sup>En todos los casos el protocolo OAI-PMH provee un encabezado con un indentificador de registro formado en base al atributo content-id y el nombre de dominio del servidor (cnx.org). Por ende se puede decir que se cuenta con dicho metadato. Ej. <identifier>oai:cnx.org:col10373</identifier>

<sup>d</sup>La fecha exportada en OAI-PMH no concuerda con ninguna de las fechas en CNXML

<sup>e</sup>El nombre viene codificado en forma de vCard

taciones similares en cuanto a que proveen acceso a un subconjunto del total de los metadatos. Sin embargo el uso del protocolo OAI-PMH brinda un mecanismo estándar de acceso y cuenta con un conjunto amplio de herramientas que lo soportan, tanto a nivel de clientes como de servidores. Esto facilitó que a lo largo de los años una gran cantidad de repositorios soportaran este protocolo<sup>38</sup>. Estas características amplían la posibilidad de reutilización de una posible solución.

En la Tabla 3.4 se presenta para cada mecanismo de acceso una síntesis de la cantidad de metadatos accesibles, la dificultad de implementación de una herramienta que lo utilice y la posibilidades de reutilización de dicha herramienta en otras fuentes de datos similares.

<sup>38</sup><http://www.openarchives.org/Register/BrowseSites>

Tabla 3.4: Resumen de los distintos métodos de acceso

Mecanismo de Acceso	Cantidad de Meta-datos	Dificultad de Implemen-tación	Posibilidad de Reutili-zación
Formato nativo / Crawler	total	media	baja
Web Service	parcial	media	baja
OAI-PMH	parcial	baja	alta

Teniendo en cuenta los patrones de publicación mencionados en 2.2.4 y el análisis de la plataforma CNX nos posiciona en un escenario donde partimos de datos estructurados en formato XML. Estos datos quedan accesibles por medio de una API REST, protocolo OAI-PMH o directamente mediante acceso HTTP al formato nativo. En la figura 3.1 se puede ver la ubicación de este caso según la naturaleza de los datos junto con las diferentes herramientas y tecnologías que se plantea usar en las distintas capas.

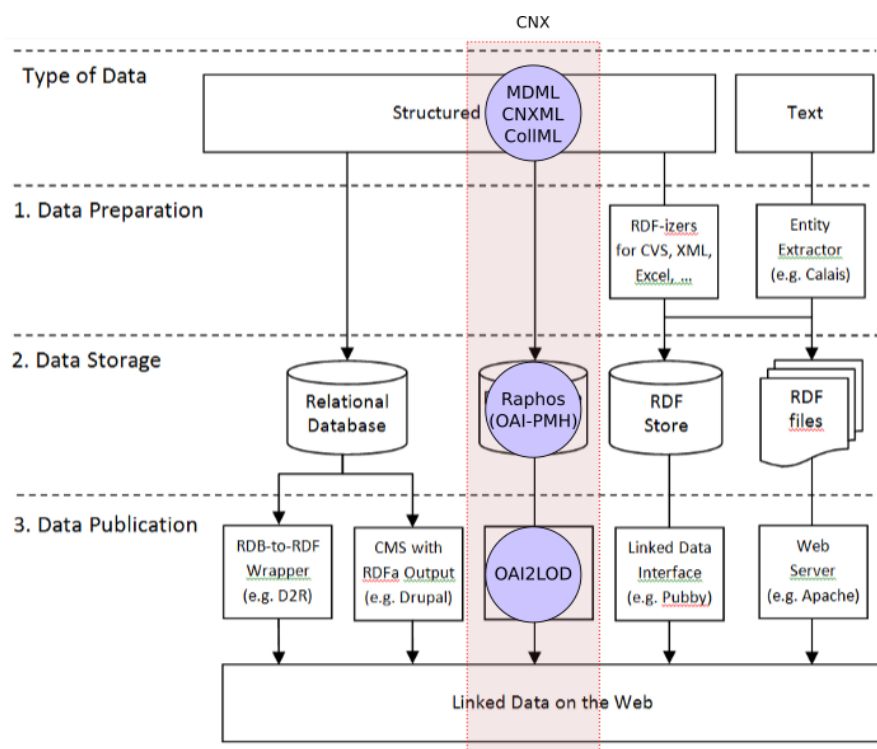


Figura 3.1: Ubicación del caso de estudio CNX.org en los Patrones de publicación

Para la publicación se propone explotar el protocolo **OAI-PMH**. Para esto se deci-

dió explorar el uso de la herramienta OAI2LOD Server [30] desarrollada para tales fines. Se propone además utilizar el formato *cnx-dc* ya que es el que brinda acceso a una mayor cantidad de metadatos.

## Licencia

Para la definición de la licencia no existen mayores controversias ya que se parte de la base que todos los contenidos están accesibles bajo licencia Creative Commons. Sin embargo, como se puede ver en la Tabla 3.5, no todos los elementos tienen el mismo sabor o versión de licencia, por lo tanto es importante preservar la licencia correcta de cada recurso.

Tabla 3.5: Cantidad de entidades que usan los distintos sabores de licencia *Creative Commons* en la plataforma

Licencia	Cantidad
<a href="http://creativecommons.org/licenses/by/3.0/">http://creativecommons.org/licenses/by/3.0/</a>	15749
<a href="http://creativecommons.org/licenses/by/4.0/">http://creativecommons.org/licenses/by/4.0/</a>	9020
<a href="http://creativecommons.org/licenses/by/2.0/">http://creativecommons.org/licenses/by/2.0/</a>	5769
<a href="http://creativecommons.org/licenses/by/1.0/">http://creativecommons.org/licenses/by/1.0/</a>	2057
<a href="http://creativecommons.org/licenses/by-nc-sa/4.0/">http://creativecommons.org/licenses/by-nc-sa/4.0/</a>	150

## Diseño de URIs

Como base para las URIs se definió usar el nombre de dominio `data.cnx.org`, que además de ser un subdominio controlado por `cnx.org`, denota claramente la intención de representar datos. A su vez se propone utilizar un prefijo `resource` y el tipo de elemento que se está describiendo, en este caso `document`, tal como propone el patrón *Patterned URIs*<sup>39</sup>. Este esquema sigue las recomendaciones propuestas en [21] donde se definen patrones diferentes para lograr una clara separación entre los datos (Abox) y las ontologías (TBox). A su vez se opta por utilizar en la medida de lo posible esquema de URIs que hagan uso de redirecciones HTTP 303, pues están recomendadas para publicar volúmenes de datos importantes<sup>40</sup>.

Para identificar a cada elemento se sigue la propuesta de la herramienta OAI2LOD, donde se utiliza como parte de la URI el identificador de cada elemento en el pro-

<sup>39</sup><http://patterns.dataincubator.org/book/patterned-uris.html>

<sup>40</sup><https://www.w3.org/TR/swbp-vocab-pub/#recipe1>

protocolo OAI-PMH. Dicho identificador es único dentro del repositorio y se compone del prefijo *oai*, el nombre de dominio y el identificador interno del elemento, por ejemplo: *oai:legacy.cnx.org:col11525*. De esta forma cada URI queda identificada por: <http://data.cnx.org/resource/document/<id-oai-pmh>>

### 3.1.2. Modelado

Para el modelo de datos se aprovechó el esquema del formato de intercambio *cnx-dc* utilizado en el protocolo OAI-PMH. Dicho formato es una extensión del formato *oai-dc*<sup>41</sup> a la que se le agregan atributos extra definidos en la plataforma CNX. Por tal motivo la mayoría de los elementos son propiedades del vocabulario *Dublin Core* de amplia difusión. Para representar los atributos extra se siguió la recomendación de reutilizar vocabularios existentes. Para esto se realizaron los siguientes mapeos:

- El tópico principal *cnxdc:cnx-subject* se mapeó a *dcterms:subject*
- Para el atributo *cnxdc:maintainer* se utilizó *dc:contributor*
- Para el vínculo entre la colección y los módulos que la componen se utilizó *dc:hasPart*

### 3.1.3. Generación de Datos

Como se mencionó anteriormente para la recolección y generación de Datos Enlazados se seleccionó la herramienta OAI2LOD Server<sup>42</sup>. El cometido de OAI2LOD es exponer cualquier repositorio de metadatos compatible con OAI-PMH como Datos Enlazados. Su arquitectura extensible permite agregar nuevos formatos de intercambio o modificar existentes por medio de transformaciones XSLT. Su front-end se basa en el Servidor D2R<sup>43</sup> y proporciona un mecanismo de dereferenciación así como una terminal SPARQL. En la figura 3.2 se muestra un diagrama de la arquitectura de la aplicación donde se resaltan los componentes que fueron extendidos en esta instancia. Por un lado el componente *OAI-PMH Harvester* que se comunica mediante el protocolo OAI-PMH con la fuente de datos, y por otro la transformación XSLT que es la encargada en última instancia de realizar la transformación a RDF.

---

<sup>41</sup>[https://www.openarchives.org/OAI/2.0/oai\\_dc.xsd](https://www.openarchives.org/OAI/2.0/oai_dc.xsd)

<sup>42</sup><https://github.com/behas/oai2lod>

<sup>43</sup><http://d2rq.org/>

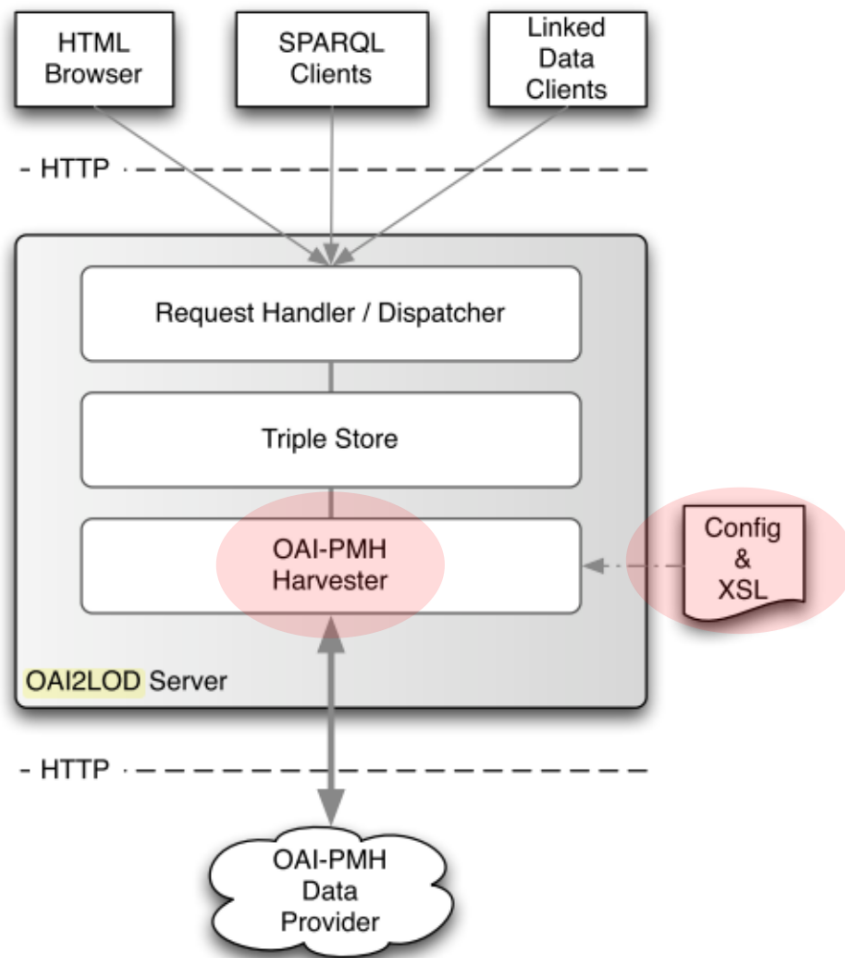


Figura 3.2: Componentes de OAI2LOD Server [31].

Para poder procesar los metadatos fue necesario crear una transformación XSLT específica, ya que se utilizó un formato de intercambio no estándar como *cnx-oai*. Dicha transformación es la encargada de convertir el XML de cada registro obtenido en su correspondiente representación RDF. En la figura 3.2 se muestra el ejemplo de un registro en formato *cnx-oai*. Allí se puede ver el encabezado del registro con el identificador y la fecha, así como la sección de metadatos que contiene la lista de campos estándar del esquema *Dublin Core* (utilizando el prefijo *cnxdc:*), y la lista de campos extra definidos por CNX (utilizando el prefijo *cnxdc:*). Por otro lado el Ejemplo A.1 del Apéndice A muestra la transformación XSLT completa creada para tales fines. Además del mapeo directo de campos, la transformación se encarga de generar la URI correcta del recurso definido en eta-

pas previas, concatenando el prefijo con el identificador del registro extraído del elemento `<identifíer>` (*oai:legacy.cnx.org:col11195* en el ejemplo). También fue necesario separar las palabras clave, ya que fueron aplanadas por el protocolo en un único campo `<dc:subject>` separado por comas.

Ejemplo 3.2: Ejemplo de un registro obtenido a través de la interfaz OAI-PMH utilizando el formato *cnx-oai*.

```

1  ...
2  <record>
3    <header>
4      <identifíer>oai:legacy.cnx.org:col11195</identifíer>
5      <datestamp>2010-03-31T03:21:21Z</datestamp>
6    </header>
7    <metadata>
8      <cnxdc:dc xmlns:cnxdc="http://cnx.org/technology/schemas/cnx_dc/"
9                xmlns:dc="http://purl.org/dc/elements/1.1/"
10               xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
11               xsi:schemaLocation="http://cnx.rice.edu/cnx_dc/
12               http://cnx.rice.edu/technology/cnx_dc/schema/xsd/1.0/cnx
13               -dc-extension.xsd">
14        <dc:title>Involvement of Banking in the Commonwealth of
15        Independent States</dc:title>
16        <dc:creator>Valentin Antonovich Kotov</dc:creator>
17        <dc:creator>Saydullo Abdullaev</dc:creator>
18        <cnxdc:maintainer>Valentin Antonovich Kotov</cnxdc:maintainer>
19        <cnxdc:maintainer>Saydullo Abdullaev</cnxdc:maintainer>
20        <cnxdc:maintainer>Andrew R. Barron</cnxdc:maintainer>
21        <cnxdc:maintainer>James Abbey</cnxdc:maintainer>
22        <cnxdc:cnx-subject>Business , Social Sciences</cnxdc:cnx-subject>
23        <dc:subject>bank , banking , business , CIS , Commonwealth of Independent
24        States , company , crisis , government , investor , share , shareholder , stock<
25        /dc:subject>
26        <dc:description></dc:description>
27        <dc:language>en</dc:language>
28        <dc:date>2010-03-31T03:21:21Z</dc:date>
29        <dc:identifíer>http://legacy.cnx.org/contents/c64e29d0-890c-40aa
30        -9cf2-f16ad8780e61@4.1</dc:identifíer>
31        <dc:rights>http://creativecommons.org/licenses/by/3.0/</dc:rights
32        >
33      </cnxdc:dc>
34    </metadata>
35  </record>
36  ...

```

En paralelo se consideró importante poder mantener la relación entre la colección y los módulos que la componen. Dicha información se pierde al utilizar el protocolo OAI-PMH, pero es accesible a través de la interfaz de servicio REST proporcionada por CNX. Por tal motivo se exploró las posibilidades de extensión

de la herramienta y se llegó a un enfoque híbrido en donde, además del protocolo OAI-PMH, se utilizan los servicios REST para complementar dicha información. Para esto se modificó el componente de cosecha de datos (*OAI-PMH Harvester* en la figura 3.2). Dicho componente está desarrollado en Java de forma tal que permite su extensión. Aprovechando este aspecto se creó la clase `OAIHarvestingJobCnx` que extiende la clase principal `OAIHarvestingJob`<sup>44</sup>. Esto permitió reutilizar prácticamente todo el código existente, por lo que únicamente fue necesario sobrescribir el constructor y la operación `harvestMetadata()`. Esta última es la operación encargada de consultar todos los elementos recuperados, identificar las colecciones y generar los enlaces. Para discriminar las colecciones de los módulos se utilizó una expresión regular sobre el identificador de cada elemento, buscando las que cumplan con el siguiente patrón `.*\/oai:legacy.cnx.org:(col\\d*)\$`. Una vez que se detecta una colección se utiliza el servicio REST comentado anteriormente para obtener los módulos de la misma: <http://legacy.cnx.org/content/<id-coleccion>/latest/containedModuleIds>. Finalmente se crea un vínculo (utilizando la propiedad *dc:hasPart*) entre la colección y el módulo, para cada uno de los módulos retornados por el servicio web. En la Figura 3.3 se puede ver un diagrama de clases simplificado de la herramienta donde se resaltaron en color verde las clases modificadas o creadas.

## Integración con otras fuentes

Para la generación de enlaces a otras fuentes se seleccionó el campo de temática principal (*Main Subject*) denotado por la propiedad *dct:subject*, y los campos de autoría denotados por las propiedades *dc:creator* y *dc:contributor*. En la generación de enlaces se usaron dos mecanismos diferentes. Por un lado se utilizaron las funcionalidades de generación de enlaces provistas por OAI2LOD para detectar enlaces entre autores de CNX y personas de DBpedia. Por otro lado se crearon enlaces de forma estática durante la generación de RDF aprovechando el hecho de que el tema principal tiene un conjunto acotado de valores.

**Tema principal (*dct:subject*).** Existen varios vocabularios controlados que pueden ser usados para organizar y catalogar recursos en la web de datos. Algunos están especializados en información bibliográfica y recursos digitales como por ejemplo el tesoro de UNESCO<sup>45</sup>, o la lista de encabezamiento de materias de la Biblioteca del Congreso de Estados Unidos (Library of Congress Subject Hea-

---

<sup>44</sup><https://github.com/behas/oai2lod/blob/master/src/at/ac/univie/mminf/oai2lod/oai/OAIHarvestingJob.java>

<sup>45</sup><http://vocabularies.unesco.org/browser/thesaurus/en>



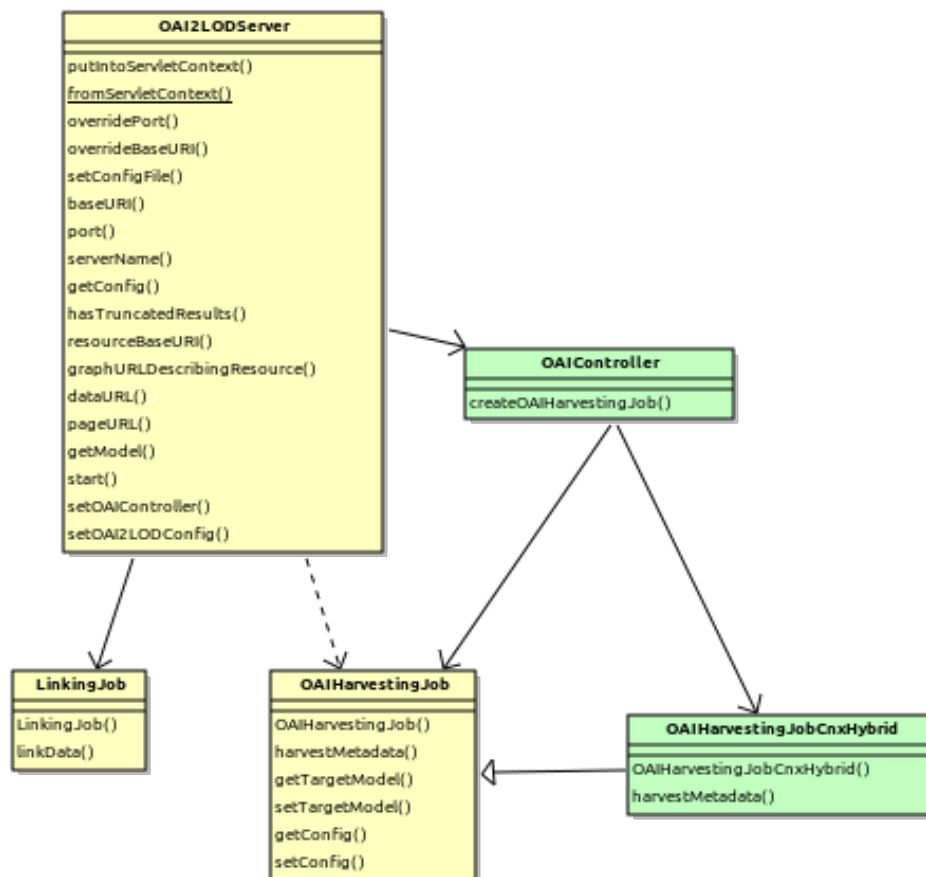


Figura 3.3: Diagrama de clases simplificado donde se muestran los componentes principales de OAI2LOD Server así como las clases que fueron creadas o modificadas (en color verde)

dings)<sup>46</sup>. Otros son vocabularios de propósito general como las categorías de clasificación propias de DBpedia<sup>47</sup>. Para esta instancia se utilizaron categorías de DBpedia ya que es una fuente de datos sumamente referenciada, lo que amplía las posibilidades de integración y reuso de los datos. En este caso se aprovechó el hecho de que el rango de posibles valores para este campo es un conjunto fijo y reducido de categorías<sup>48</sup>. Esto hace viable que se pueda seleccionar manualmente una categoría equivalente en el vocabulario objetivo. En la tabla 3.6 se puede ver la

<sup>46</sup><http://id.loc.gov/>

<sup>47</sup>[http://dbpedia.org/page/Category:Main\\_topic\\_classifications](http://dbpedia.org/page/Category:Main_topic_classifications)

<sup>48</sup><http://legacy.cnx.org/help/reference/subjects>

Tabla 3.6: Cantidad de documentos en cada tema principal con su respectiva categoría DBpedia vinculada

Tema principal CNX	Cantidad	Categoría DBpedia
Science and Technology	14976	<i>dbc:Science, dbc:Technology</i>
Mathematics and Statistics	6808	<i>dbc:Mathematics, dbc:Statistics</i>
Social Sciences	4164	<i>dbc:Social_sciences</i>
Humanities	2691	<i>dbc:Humanities</i>
Arts	1590	<i>dbc:Arts</i>
Business	898	<i>dbc:Business</i>
Test/Draft	136	-
OpenStax Featured	27	-
CNX Featured	13	-

lista de temas principales disponibles en CNX junto con la cantidad de contenidos y la categoría de DBpedia seleccionada para el vínculo. El enlace se generó directamente durante el proceso de generación de RDF utilizando la transformación XSLT como se puede ver en la plantilla XSLT *splitSubject* a partir de la línea 189 en el Ejemplo A.1.

**Autores (dc:creator, dc:contributor).** Para el caso de las propiedades relacionadas con la autoría de los documentos se buscó generar enlaces a entidades de tipo persona en DBpedia<sup>49</sup>. La generación de enlaces se realizó utilizando el componente de detección de enlaces provisto por la herramienta. Dicho componente se configuró para comparar la similitud entre el nombre de autor de CNX y el nombre de la persona en DBpedia utilizando el algoritmo de comparación de cadenas de caracteres *Levenshtein*. A pesar de utilizar umbrales altos de exigencia en la exactitud de comparación las pruebas exploratorias iniciales fueron costosas y arrojaron un número muy elevado de falsos positivos. Esto sucede ya que la cantidad de pares de objetos crece de forma cuadrática con el tamaño del conjunto de datos. Por consiguiente el cálculo de similitud entre todos los pares de entidades se vuelve poco práctico y particularmente costoso en grandes volúmenes de datos o funciones de similitud complejas. Por tal motivo las pruebas exploratorias decantaron de forma natural en la búsqueda de métodos de bloqueo que puedan ser aplicables a este contexto.

Los métodos de bloqueo (o *blocking*) reducen este problema mediante la selección eficiente de pares de objetos aproximadamente similares para los cálculos de distancia subsiguientes, dejando fuera los pares restantes que son considerados diferentes

<sup>49</sup><http://downloads.dbpedia.org/current/core/>

Tabla 3.7: Autores identificados en recursos de tipo Persona en DBpedia y la cantidad de publicaciones de CNX en los que figura como autor.

Entidad DBpedia	Cantidad de publicaciones
<i>dbr:Andreas_Luttge</i>	1
<i>dbr:C._Sidney_Burrus</i>	708
<i>dbr:Lydia_Kavraki</i>	39
<i>dbr:Mark_Embree</i>	4
<i>dbr:Matthias_Felleisen</i>	38
<i>dbr:Moshe_Vardi</i>	48
<i>dbr:Richard_Baraniuk</i>	784

[32]. El bloqueo intenta restringir las comparaciones solo a aquellas entidades para las cuales coinciden uno o más atributos particularmente discriminatorios. Es importante mencionar que este tipo de técnicas tienen el efecto de aumentar el valor predictivo positivo (**precisión**) a expensas de la **sensibilidad** (o *recall*).

Siguiendo esta línea, con el objetivo de minimizar los errores y reducir el tiempo de cómputo, se intentó reducir el universo de comparación buscando un subconjunto más específico de personas de DBpedia. Esto se logró filtrando las personas catalogadas como Científicos (23355 recursos de tipo *dbo:Scientist*), que a su vez coexistieron en el tiempo con la plataforma CNX, es decir, que están vivos o murieron luego de 2003 (aproximadamente 6000 recursos que cumplen ambas condiciones). Luego de ajustes de configuración y reducción del universo de comparación el proceso de detección arrojó únicamente 7 enlaces correctos de 22 detectados (15 falsos positivos) en un total de 2132 autores de CNX. El vínculo se representó con la propiedad *dct:creator*. Es interesante destacar que todos los casos detectados en DBpedia tienen alguna relación con la universidad Rice, expresada mediante los conceptos *dbc:Rice\_University\_faculty* *dbc:Rice\_University\_alumni*. La Tabla 3.7 muestra un resumen de los recursos identificados en DBpedia junto con la cantidad de publicaciones de CNX en las que figuran como autores.

Cabe mencionar que la detección de falsos positivos no es tarea fácil cuando se dispone de pocos datos para validar la identidad de las entidades que se comparan, en este caso Personas. A esto se le suma el hecho que en ambos conjuntos de datos se tiene una gran cantidad de individuos, lo que hace inviable un procesamiento manual completo. Por tal motivo se hizo una comprobación manual aleatoria de diferentes muestras. Por ejemplo priorizando los casos de detección de múltiples enlaces desde la misma entidad CNX, donde es muy probable que haya falsos positivos. Otra estrategia fue utilizar datos de las entidades de DBpedia enlazadas,

buscando características que den indicios de que se trata de un falso positivo. Por ejemplo, la fecha de nacimiento y muerte.

### 3.1.4. Publicación

El uso de OAI2LD Server resuelve varios aspectos de la etapa de publicación utilizando los servicios provistos por D2RQ<sup>50</sup>. Este último es el que se encarga de la negociación de contenido, la dereferenciación y de exponer un terminal SPARQL. En la imagen 3.4 se muestra un ejemplo de página web desplegada al acceder a un recurso publicado desde un navegador. En lo concerniente a la descripción de la fuente de datos. Si bien D2RQ tiene soporte para agregar metadatos mediante el uso de VOID y Provenance Vocabulary<sup>51</sup> dichas configuraciones no fueron expuestas como parte de la configuración de OAI2LOD Server y por ende no se pueden establecer.

---

<sup>50</sup><http://d2rq.org/>

<sup>51</sup><http://d2rq.org/d2r-server#metadata-templates-default>

Property	Value
dc:contributor	David Marasco
dc:creator	David Marasco
dc:date	2017-02-10T20:05:32Z
dc:description	This introductory, algebra-based, one-quarter college physics book is grounded with real-world examples, illustrations, and explanations to help students grasp key, fundamental physics concepts. This online, fully editable and customizable title includes learning objectives, concept questions, links to labs and simulations, and ample practice opportunities to solve traditional physics application problems.
dcterms:hasPart	<http://localhost:2020/resource/item/oai:legacy.cnx.org:m42185>
dcterms:hasPart	<http://localhost:2020/resource/item/oai:legacy.cnx.org:m42186>
dcterms:hasPart	<http://localhost:2020/resource/item/oai:legacy.cnx.org:m42187>
dcterms:hasPart	<http://localhost:2020/resource/item/oai:legacy.cnx.org:m42189>
dcterms:hasPart	<http://localhost:2020/resource/item/oai:legacy.cnx.org:m42193>
dcterms:hasPart	<http://localhost:2020/resource/item/oai:legacy.cnx.org:m42195>
dcterms:hasPart	<http://localhost:2020/resource/item/oai:legacy.cnx.org:m42196>
dcterms:hasPart	<http://localhost:2020/resource/item/oai:legacy.cnx.org:m42204>
dcterms:hasPart	<http://localhost:2020/resource/item/oai:legacy.cnx.org:m42205>
dcterms:hasPart	<http://localhost:2020/resource/item/oai:legacy.cnx.org:m42206>
dcterms:hasPart	<http://localhost:2020/resource/item/oai:legacy.cnx.org:m42208>
dc:identifier	<http://legacy.cnx.org/contents/2bc4ce2d-87ed-4e89-a47e-6c8b9673017a@1.3>
dc:language	en
oai:origin	<http://cnx.org/content/OAI?verb=GetRecord&metadataPrefix=cnx_dc&identifier=oai:legacy.cnx.org:col12032>
dc:rights	http://creativecommons.org/licenses/by/4.0/
dc:subject	Ohm's Law
dc:subject	ac circuits
dc:subject	circuits
dc:subject	college physics
dc:subject	dc instruments
dc:subject	electric charge and electric field
dc:subject	electric current
dc:subject	electric potential
dc:subject	electrical technologies
dc:subject	electromagnetic induction
dc:subject	electromagnetic waves
dc:subject	energy
dc:subject	fluid dynamics
dc:subject	fluid statics
dc:subject	gas laws
dc:subject	heat and transfer methods
dc:subject	kinetic theory
dc:subject	magnetism
dc:subject	resistance
dc:subject	temperature
dc:subject	thermodynamics
dcterms:subject	<dbc:Mathematics>
dcterms:subject	<dbc:Science>
dcterms:subject	<dbc:Statistics>
dcterms:subject	<dbc:Technology>
dc:title	OpenSTAX Text for Physics 2B - Marasco - Chapter 11 (11+12)
rdf:type	oai:item

Generated by [OAI Server](#)

Figura 3.4: Ejemplo de visualización de una colección desde un navegador web.

## 3.2. Producción bibliográfica del InCo

Como se mencionó anteriormente el objetivo es poder analizar la producción bibliográfica de los docentes del Instituto de Computación. En este aspecto una iniciativa destacada es el proyecto Colibri<sup>52</sup>, que oficia de repositorio institucional para la Universidad de la República. Lamentablemente al día de hoy no dispone de un mecanismo de acceso a los datos en bruto, es decir, no es posible descargar el conjunto completo y primario de los datos almacenados en un formato que facilite el procesamiento por computadoras. A esto se le suma el hecho de que una gran cantidad de docentes aun no carga sus publicaciones en la plataforma.

En este contexto una de las mayores dificultades encontradas es no poder acceder a una fuente de datos única y completa que esté disponible públicamente. Muchas veces dicha información se encuentra dispersa, incompleta y en distintos formatos, lo cual dificulta el acceso y procesamiento de los datos. Algunos investigadores cargan sus publicaciones en su página web personal usando formatos poco estructurados<sup>53 54</sup>. Otros optan por utilizar las funcionalidades que provee el sitio web institucional<sup>55</sup>. Y en algunos casos la información se publica en múltiples sitios<sup>56 57</sup> lo que en ocasiones genera discrepancias si no se mantienen sincronizados.

Por otro lado existen varios sitios web que permiten acceder a información bibliográfica en línea. Algunos son concentradores (*hubs*) que recolectan información mediante procesos de crawling sobre sitios web de editoriales, conferencias y revistas (*CiteSeer*<sup>58</sup>, *DBLP*<sup>59</sup>, *Semantic Scholar*<sup>60</sup>, entre otros). Otros dan la posibilidad de registrarse y cargar un perfil personal con el fin de generar comunidades en torno a sus áreas de investigación (*Google Scholar*<sup>61</sup>, *Research Gate*<sup>62</sup>, entre otros). Si bien estos sitios concentran mucha información, en general proporcionan funcionalidades limitadas de análisis sobre los datos. Algunos soportan APIs con acceso restringido, otros requieren una solicitud explícita de permiso para acceder a los datos (o a un subconjunto de ellos). A esto se suma que muchas veces existen limitaciones de uso de determinados datos a consecuencia del licenciamiento y

---

<sup>52</sup><https://www.colibri.udelar.edu.uy/>

<sup>53</sup><https://www.fing.edu.uy/~mmartine/>

<sup>54</sup><https://www.fing.edu.uy/~pardo/>

<sup>55</sup><https://www.fing.edu.uy/inco/institucional/docentes/lorenae>

<sup>56</sup><https://www.fing.edu.uy/inco/institucional/docentes/sergion>

<sup>57</sup><https://www.fing.edu.uy/~sergion>

<sup>58</sup><http://citeseerx.ist.psu.edu/>

<sup>59</sup><https://dblp.uni-trier.de/>

<sup>60</sup><https://www.semanticscholar.org/>

<sup>61</sup><https://scholar.google.com.uy/>

<sup>62</sup><https://www.researchgate.net/>

los derechos que poseen ciertas editoriales sobre las publicaciones. En general son pocos los que permiten el acceso a los datos en crudo para poder procesarlos y analizarlos por cuenta propia.

En el escenario planteado se desea recopilar la información bibliográfica de los docentes del InCo. Para esto se propone utilizar tres fuentes de datos diferentes: la lista de docentes publicada en el sitio web del InCo, la información bibliográfica cargada en el sitio web de la Facultad de Ingeniería, y los datos de DBLP actualmente disponibles en forma de Datos Enlazados.

### 3.2.1. Especificación

La lista de docentes se encuentra publicada en forma de tabla **HTML** en la sección del InCo del sitio web de la Facultad de Ingeniería<sup>63</sup>. La tabla contiene los siguientes datos: grado, cantidad de horas, si trabaja en régimen de dedicación total, el número de oficina y el teléfono interno. Además se presenta un enlace a la página personal del docente. Dicho enlace contiene como parte de la URL el nombre de usuario que es único dentro de la institución. Nos referiremos a esta fuente como *Docentes InCo*.

En el sitio web de la Facultad de Ingeniería también se puede encontrar un catálogo de publicaciones<sup>64</sup> cargadas por los docentes de la institución. El sitio web está desarrollado utilizando el gestor de contenidos *Drupal*<sup>65</sup>, y cuenta con un módulo para la gestión de datos bibliográficos denominado *Bibliography*<sup>66</sup>. Dicho módulo le permite a los docentes cargar las referencias bibliográficas de sus publicaciones y presentarlas de forma amigable en sus páginas personales. A su vez da la posibilidad de realizar búsquedas sobre toda la base de publicaciones y autores. Un aspecto importante es que permite descargar un archivo completo (o *dump*) de toda la información bibliográfica almacenada en formato **BibTex**<sup>67</sup>, lo cual es sumamente útil para su reutilización. Nos referiremos a esta fuente como *Biblio FIng*.

Como se puede observar en ambos casos partimos de datos representados en formatos semiestructurados (HTML y BibTex). En la figura 3.5 se puede ver la ubicación de este caso según la naturaleza de los datos junto con las diferentes herramientas y tecnologías que se planea utilizar en las distintas capas. Para obtener las

---

<sup>63</sup><https://www.fing.edu.uy/inco/institucional/docentes>

<sup>64</sup><https://www.fing.edu.uy/publicaciones>

<sup>65</sup><http://drupal.org>

<sup>66</sup><https://www.drupal.org/project/biblio>

<sup>67</sup><http://www.bibtex.org/>

publicaciones de los docentes se utilizaron las publicaciones disponibles en DBLP. Para esto se aprovechó la fuente de datos *DBLP++*<sup>68</sup> publicada en el contexto del proyecto *FacetedDBLP*<sup>69</sup> del *L3S Research Center*. Esta fuente de datos es construida a partir de los datos publicados periódicamente por DBLP, sumado a palabras clave y resúmenes adicionales obtenidos de páginas web públicas. Una gran ventaja es que este proyecto ya cuenta con los datos disponibles en forma de Datos Enlazados, provee una terminal SPARQL así como un *dump* para descargar el conjunto completo de datos en RDF. Para el presente estudio se utilizaron datos de octubre de 2017.

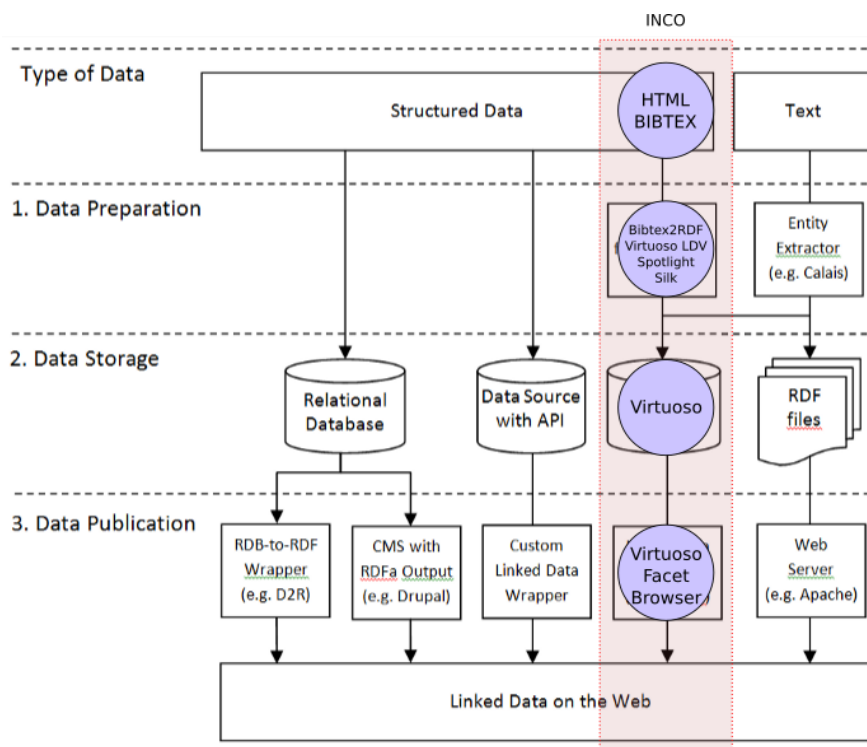


Figura 3.5: Ubicación del escenario teniendo en cuenta la naturaleza de los datos

## Licenciamiento

Cuando se evalúa la publicación de datos sea cual sea el formato, se debe prestar especial atención a la normativa vigente relacionada con la protección de datos personales y derecho de autor. Si bien muchas de estas normativas están alineadas con acuerdos internacionales, suelen variar de país en país.

<sup>68</sup><https://dblp.l3s.de/dblp++.php>

<sup>69</sup><https://dblp.l3s.de>



En el presente estudio se manejan datos que actualmente se encuentran disponibles públicamente en la web. En el caso de los docentes se publican datos básicos de contacto institucional, así como algunos datos referentes al tipo de relación laboral que poseen con una institución estatal, como por ejemplo, la cantidad de horas que trabajan o el grado que ocupan. En ningún momento se manejan datos que puedan infringir la Ley de protección de datos personales de Uruguay (Ley N° 18331<sup>70</sup>). Por lo anteriormente dicho se considera posible la publicación de estos datos así como la utilización de una licencia que permita el libre acceso y uso de los mismos, como por ejemplo **CC-BY**<sup>71</sup> o la licencia de datos abiertos de gobierno sugerida por AGESIC<sup>72</sup>.

Para las publicaciones se manejan únicamente los metadatos que las describen y no la obra en sí misma, por ende no hay riesgo de infringir ninguna restricción de *copyright*. En lo concerniente a la licencia de las publicaciones se mantiene la licencia correspondiente en caso de que exista el dato.

## Diseño de URIs

De forma similar al caso anterior se propone como dominio base el uso del subdominio `data.fing.edu.uy`. A su vez se utilizan los términos **resource** y **schema** como parte del patrón de **URI**, de esta forma se mantiene una separación clara entre los datos (Abox) y los vocabularios (TBox) respectivamente.

Para las publicaciones de la base bibliográfica se utilizó el término **biblio** como parte del patrón para la URI. Para identificar las entidades bibliográficas se utilizó el identificador único asignado por el módulo de gestión bibliográfica, obteniéndose URIs del siguiente estilo: <http://data.fing.edu.uy/resource/biblio/<bibtext-id>>.

Para el caso de los docentes no disponemos de atributos publicados de forma explícita que puedan ser utilizados para su identificación. Si bien se cuenta con el nombre propio, este dato no garantiza unicidad y podría presentar conflictos en docentes con el mismo nombre. Un dato interesante es el nombre de usuario que se codifica como parte de la URL para acceder a la página personal del docente. Como se puede ver en el ejemplo del docente *Alverto Pardo*, su página personal en sitio web del InCo es <http://www.fing.edu.uy/inco/institucional/docentes/pardo> don-

---

<sup>70</sup><https://www.impo.com.uy/bases/leyes/18331-2008>

<sup>71</sup><https://creativecommons.org/licenses/by/4.0/>

<sup>72</sup><https://www.agesic.gub.uy/innovaportal/file/6327/1/licencia-de-datos-abiertos.pdf>

de el último segmento de la URL (**pardo**) es el nombre de usuario y coincide con la página personal a nivel institucional <http://www.fing.edu.uy/~pardo>. De esta forma se obtienen URIs que siguen el siguiente patrón <http://data.fing.edu.uy/resource/docente/<nombre-usuario>>. En el ejemplo anterior la URI resultado es la siguiente <http://data.fing.edu.uy/resource/docente/pardo>.

Una alternativa posible es utilizar el identificador creado en el contexto del control de autoridades de alguna organización reconocida. Originalmente se llamó control de autoridades a la práctica de instaurar y mantener el control de los materiales bibliográficos de un catálogo, por medio de la utilización de convenciones, categorías o identificadores numéricos para cada elemento con el que los catalogadores puedan distinguir entre nombres semejantes o idénticos. Con la aparición de Internet la necesidad de interactuar entre catálogos de distintas organizaciones resaltó el valor de contar con prácticas y mecanismos de catalogación estándar. Poco a poco los distintos catálogos comenzaron a establecer consorcios cooperativos a nivel regional y mundial, como **Online Computer Library Center (OCLC)**<sup>73</sup>, contribuyendo a la creación de registros y estándares a nivel mundial para la identificación de todo tipo de elementos. Hoy en día existen varios ejemplos destacados de estándares para la identificación, tales como **Open Researcher and Contributor ID (ORCID)**<sup>74</sup> para autores científicos o académicos, o **Digital Object Identifier (DOI)**<sup>75</sup> para publicaciones digitales. Contar con este tipo de identificadores es sumamente útil ya que permiten identificar unívocamente distintas entidades y resolver posibles casos de ambigüedad. Tal es el punto que al día de hoy existen muchas revistas científicas, fuentes de datos bibliográficos, y plataformas académicas que permiten cargar y utilizar el identificador ORCID para la identificación de los investigadores. En lo concerniente a este caso cabe destacar que desde el año 2017 DBLP ha intensificado sus esfuerzos para vincular sus perfiles de autores con sus identificadores ORCID. La información de ORCID ahora se agrega regularmente al conjunto de datos DBLP<sup>76</sup>.

Para el caso de los Docentes del InCo se intentó obtener los identificadores ORCID de forma automática. Para esto se utilizó la herramienta *OpenRefine*<sup>77</sup> y el servicio de reconciliación *Conciliator*<sup>78</sup> a partir del nombre del docente. Lamentablemente fueron pocos los identificadores encontrados con una posibilidad clara de ser utilizados. Gran cantidad de los perfiles ORCID encontrados no contaban con datos

---

<sup>73</sup><https://www.oclc.org>

<sup>74</sup><https://orcid.org/>

<sup>75</sup><http://doi.org>

<sup>76</sup><http://dblp.org/news/2018#news-17958202>

<sup>77</sup><http://openrefine.org>

<sup>78</sup><https://github.com/codeforkjeff/conciliator>

Tabla 3.8: Mapeo de los datos de Docentes disponibles propiedades de distintos vocabularios.

Metadato	Propiedad	Ejemplo
tipo	rdfs:subClassOf	foaf:Person
nombre	foaf:name	Laura Gonzalez
grado	fing:grado	3
horas	fing:horas	30
dedicación total	fing:dedicacionTotal	true
oficina	fing:oficina	8
teléfono	foaf:phone	+598 27114244 1008
página web	foaf:homepage	<a href="http://www.fing.edu.uy/...">http://www.fing.edu.uy/...</a> <sup>79</sup>
ORCID	dbo:orcidId	-

suficientes que permitieran desambiguar o asegurar la identidad de los docentes. En este sentido es deseable que los propios docentes tengan la oportunidad de identificarse proporcionando dicho identificador con parte de sus datos de perfil institucionales.

### 3.2.2. Modelado

Para representar a los docentes de la fuente *Docentes InCo* se utilizaron mayoritariamente propiedades del vocabulario FOAF. Para los atributos que describen aspectos de la relación laboral con la institución, tales como el grado del docente, la cantidad de horas, o si trabaja en régimen de dedicación total, fue necesario crear un nuevo vocabulario ya que no se encontraron propiedades equivalentes en vocabularios existentes. La tabla 3.8 muestra un mapeo de las distintas columnas de la tabla de docentes, la propiedad y vocabulario asignado y un ejemplo de datos. En el Apéndice B se puede consultar el vocabulario completo, mientras que en la Figura 3.6 se puede ver una representación gráfica simplificada del mismo.

En el caso de los datos de *Biblio FIng* se aprovechó el hecho de que los datos ya se encuentran representados en un formato ampliamente difundido como lo es *BibTex*. Esto incrementa las probabilidades de encontrar vocabularios que lo describan así como herramientas que faciliten su manipulación. Los vocabularios utilizados en este caso fueron los siguientes:

<sup>79</sup><https://www.fing.edu.uy/inco/institucional/docentes/lauragon>

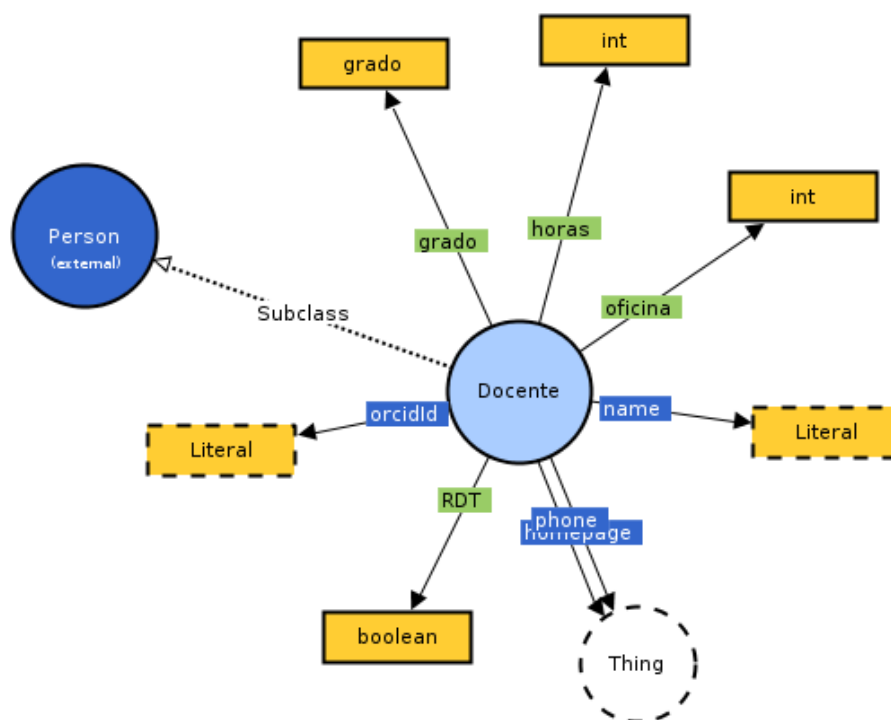


Figura 3.6: Representación gráfica del vocabulario Docentes. Las propiedades en color verde fueron definidas en el vocabulario, mientras que las azules se reutilizaron de vocabularios existentes.

- Dublin Core estándar<sup>80</sup> para metadatos bibliográficos básicos.
- Dublin Core Metadata Terms<sup>81</sup> para refinar algunos elementos
- vCard<sup>82</sup> para la representación de datos acerca de direcciones y personas, y finalmente
- BibTex<sup>83</sup> para el resto de los elementos así como los distintos tipos de entidades (*Article*, *Book*, *Proceedings*, etc). En la Figura 3.7 se puede ver una representación gráfica de este vocabulario.

<sup>80</sup><http://purl.org/dc/elements/1.1/>

<sup>81</sup><http://purl.org/dc/terms/>

<sup>82</sup><http://www.w3.org/2001/vcard-rdf/3.0#>

<sup>83</sup><http://purl.org/net/nknouf/ns/bibtex#>

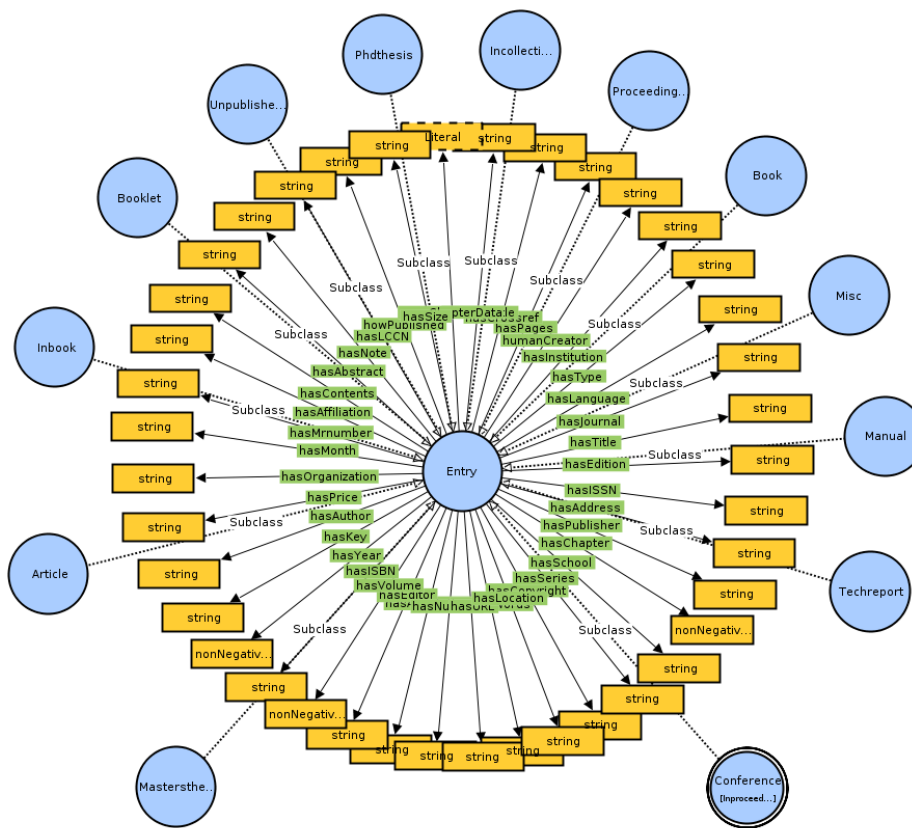


Figura 3.7: Representación gráfica del vocabulario bibtex utilizado: <http://purl.org/net/nknouf/ns/bibtex#>.

### 3.2.3. Generación

Existe una gran cantidad y variedad de herramientas que facilitan la transformación de todo tipo de formatos a **RDF**. En este caso partimos de datos en formato HTML y BibTex. En este sentido un buen punto de partida para buscar este tipo de herramientas puede ser la propia comunidad W3C<sup>84</sup>. Por otro lado para dar soporte a varias de las actividades de publicación y resolver requerimientos de almacenamiento y consulta de RDF, se utilizó la herramienta *OpenLink Virtuoso*<sup>85</sup>. Esta herramienta es la versión de código abierto del sistema *Virtuoso Universal Server*. Este sistema se puede ver como un híbrido de middleware y motor de base de datos que combina la funcionalidad de un **RDBMS** tradicional, **ORDBMS**, base

<sup>84</sup><https://www.w3.org/wiki/ConverterToRdf>

<sup>85</sup><https://virtuoso.openlinksw.com/>

de datos virtual, RDF, XML, texto libre, servidor de aplicaciones web y de archivos en un solo sistema. Es una herramienta extremadamente flexible que puede dar apoyo a varias de las etapas del ciclo de vida de los Datos Enlazados.

Para la generación del RDF correspondiente a los docentes se utilizó el componente denominado *Linked Data Views*<sup>86</sup> que forma parte de Virtuoso. Esta herramienta permite generar datos en RDF en base a una serie de correspondencias previamente definidas entre atributos de una base de datos relacional y propiedades de un vocabulario determinado. Dichas correspondencias se almacenan usando el lenguaje estándar de mapeo entre bases de datos relacionales y RDF denominado R2RML<sup>87</sup>. Una vez definidas las correspondencias entre los campos el componente se comporta similar a una vista de una base relacional, actualizando los datos RDF en la medida que cambian los datos originales de la tabla en la que se basa. Dicha tabla origen se pobló por medio de un proceso de extracción, limpieza y carga de los datos tabulares disponibles en la página de docentes del InCo utilizando la herramienta *Open Refine*<sup>88</sup>.

Los datos bibliográficos se descargaron en formato BibTex de la página institucional y se transformaron a RDF mediante la herramienta *bibtex2rdf*<sup>89</sup>. Esta aplicación permite procesar archivos en formato BibTex y generar un archivo equivalente en RDF. Una característica no menor de esta herramienta es que permite configurar varios aspectos de la transformación, como por ejemplo los vocabularios y propiedades que se utilizarán para los distintos elementos de BibText. Este aspecto es importante ya que el proceso requirió ajustes para incluir propiedades que no estaban mapeadas en la configuración por defecto de la herramienta, tales como las palabras clave o los resúmenes (abstracts) de las publicaciones.

En la Figura 3.8 se presenta un diagrama simplificado del proceso de extracción, transformación y carga de los datos, donde se puede ver el flujo de información y las herramientas más importantes utilizadas. Para almacenar el RDF resultante se definieron tres conjuntos de grafos diferentes: 1) grafos temporales, donde se almacenan datos que son producto de aplicar transformaciones o procesos intermedios que no deben quedar disponibles públicamente. 2) grafos con datos que son resultado final de esas operaciones intermedias, estos son los datos a publicar. 3) grafos con los enlaces a otras fuentes. Esta separación permite manipular y gestionar los distintos conjuntos de tripletas de forma más fácil e independiente.

---

<sup>86</sup><http://vos.openlinksw.com/owiki/wiki/VOS/VOSSQL2RDF>

<sup>87</sup><https://www.w3.org/TR/r2rml/>

<sup>88</sup><http://openrefine.org>

<sup>89</sup><http://www.l3s.de/~siberski/bibtex2rdf/>

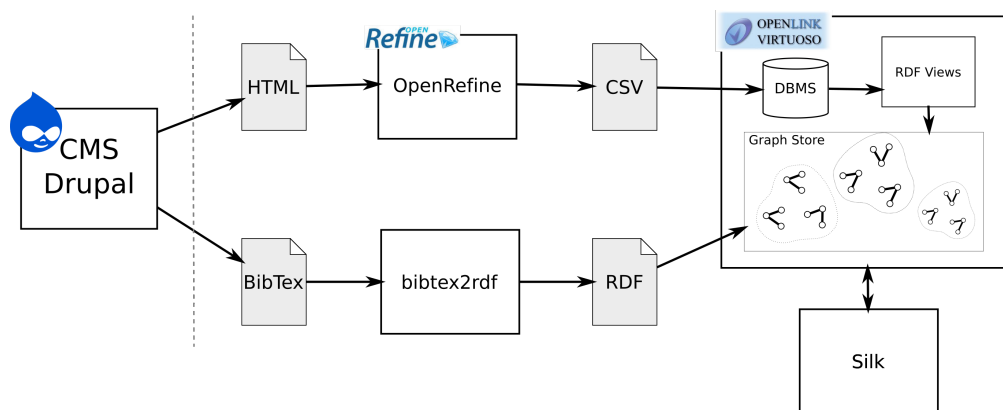


Figura 3.8: Diagrama de extracción transformación y carga de RDF

### Generación de enlaces a fuentes externas

En este caso el foco está en poder identificar la producción bibliográfica de los docentes. Por tal motivo se trabajó en la detección de enlaces entre la base de *Docentes InCo*, la fuente *Biblio FIng* y la fuente DBLP. Para la detección de enlaces se utilizó la herramienta *Silk*<sup>90</sup> y se definieron como fuentes de datos los juegos de datos publicados en la sección previa y un *dump* de la base DBLP del 22 de setiembre de 2017<sup>91</sup>. Todas estas fuentes son accedidas a través del terminal SPARQL de una instalación local de *Virtuoso*. Como métrica de comparación de nombres se utilizó n-gramas de largo 3 (tri-gramas). En [33] se puede encontrar un estudio comparativo donde se destaca como uno de los algoritmos de mejor desempeño para comparar nombres en idioma español. Previo a la comparación se aplicaron filtros de normalización de cadena de caracteres para suavizar posibles diferencias con mayúsculas, minúsculas o caracteres especiales.

La validación de los enlaces detectados se realizó de forma manual, ya que el número de docentes y publicaciones era bastante acotado. Para la evaluación de los enlaces se usó la herramienta *Silk Workbench*, que provee funcionalidades que facilitan la revisión rápida del resultado de la regla de comparación, permitiendo hacer ajustes a la heurística utilizada. Para los casos de ambigüedad más complejos se consultó información en otras fuentes de datos web, tales como las páginas personales de los docentes o los perfiles accesibles desde el **Sistema Nacional de Investigadores (SNI)**<sup>92</sup>. En la Figura 3.9 se puede ver un ejemplo de revisión del enlace detectado entre una publicación de *Biblio FIng* y una de DBLP.

<sup>90</sup><http://silkframework.org/>

<sup>91</sup><http://dblp.l3s.de/dblp++.php>

<sup>92</sup><http://sni.org.uy/>

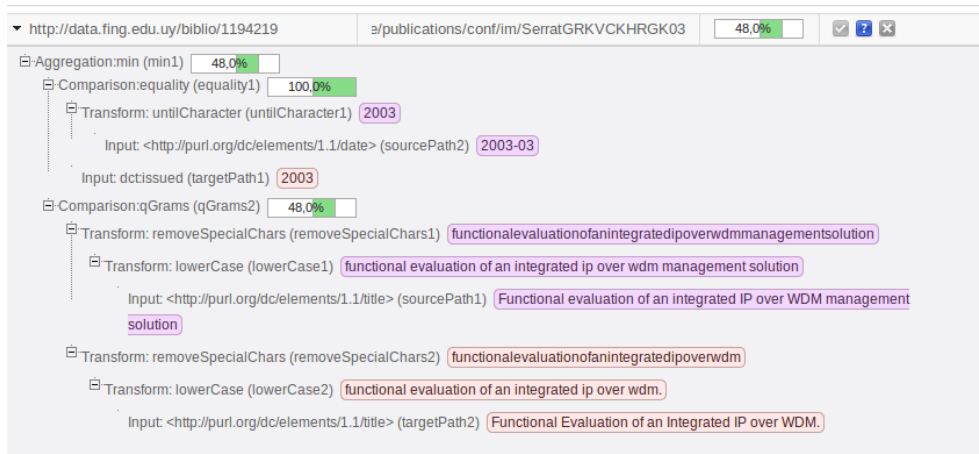


Figura 3.9: Ejemplo de revisión y validación de enlaces detectados entre publicaciones de *Biblio FIng* y DBLP utilizando la herramienta *Silk Workbench*.

En la Figura 3.10 se puede ver un diagrama de las distintas fuentes con su población de entidades de interés, así como la cantidad de enlaces distintos que se establecieron. Cada enlace entre los conceptos de ambas fuentes se representó con la propiedad *owl:sameAs*. Los conjuntos de enlaces resultantes de cada proceso se cargaron en un grafo independiente, lo que nos permite hacer un manejo diferenciado de los datos generados. Esto es recomendable ya que muchas veces el conjunto de enlaces tiene distintos niveles de calidad o frecuencia de actualización, el tenerlos separados no solo facilita su manipulación sino que permite asociar información de procedencia a cada conjunto de datos. En este aspecto es interesante hacer mención al sistema de mantenimiento, actualización y gestión de calidad de la base de enlaces de DBpedia[34]. Esta guía<sup>93</sup> define un ciclo de vida aparte para gestionar la gran cantidad y dinamismo de las contribuciones de la comunidad a la base de enlaces de DBpedia. Entre otras cosas indica los artefactos necesarios según la modalidad de contribución que se elija, así como las convenciones de nomenclatura y uso de metadatos para describir los conjuntos de enlaces. En la Figura 3.11 se puede ver el diagrama de ciclo de vida de los enlaces mencionado anteriormente.

## Docentes InCo a autores de Biblio FIng

La generación de enlaces entre estas dos fuentes se realizó comparando directamente los nombres de las instancias de tipo *foaf:Person* de ambas bases de datos.

<sup>93</sup><https://github.com/dbpedia/links/wiki/How-To-Contribute-Links-to-DBpedia>



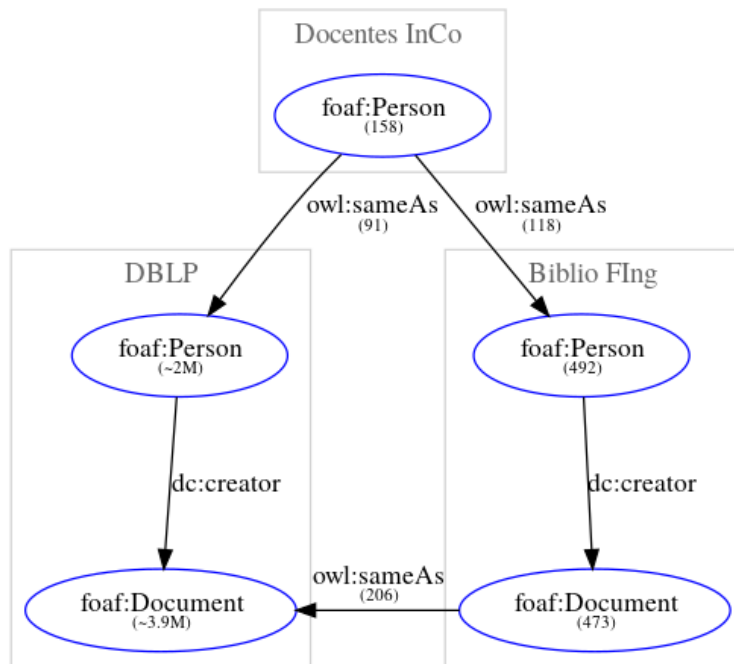


Figura 3.10: Enlaces entre los tipos de entidades de las distintas fuentes. Se muestra entre paréntesis la cantidad de tripletas.

Para esto se compararon los atributos *foaf:name* de la fuente Docentes InCo con *vcard:FN* de la fuente *Biblio FIng*, utilizando la métrica de comparación de nombre comentada previamente. En este caso se detectaron 91 enlaces correspondientes a 76 docentes diferentes en la fuente *Docentes FIng*. En la Tabla 3.9 se puede ver un resumen de cantidad y tipo de los distintos conceptos de la fuente Biblio FIng junto con la cantidad de enlaces (directos o indirectos) detectados desde Docentes del InCo. Allí se puede ver por ejemplo, que el 18% de los autores (*foaf:Person*) de la fuente *Biblio FIng* fueron enlazados con docentes del InCo (*foaf:Person*), y que a su vez son los autores de 85 de los 137 artículos (*bibtex:Articles*) de la fuente *Biblio FIng*.

### Publicaciones de BiblioFIng a publicaciones de DBLP

Para la detección de enlaces entre las publicaciones de Biblio y DBLP se comparó el título y año de cada publicación. En este caso se compararon las instancias de tipo *foaf:Document*. Para los títulos se compararon los atributos *dc:title* de ambas fuentes utilizando n-gramas luego de normalizar las cadenas de caracteres.

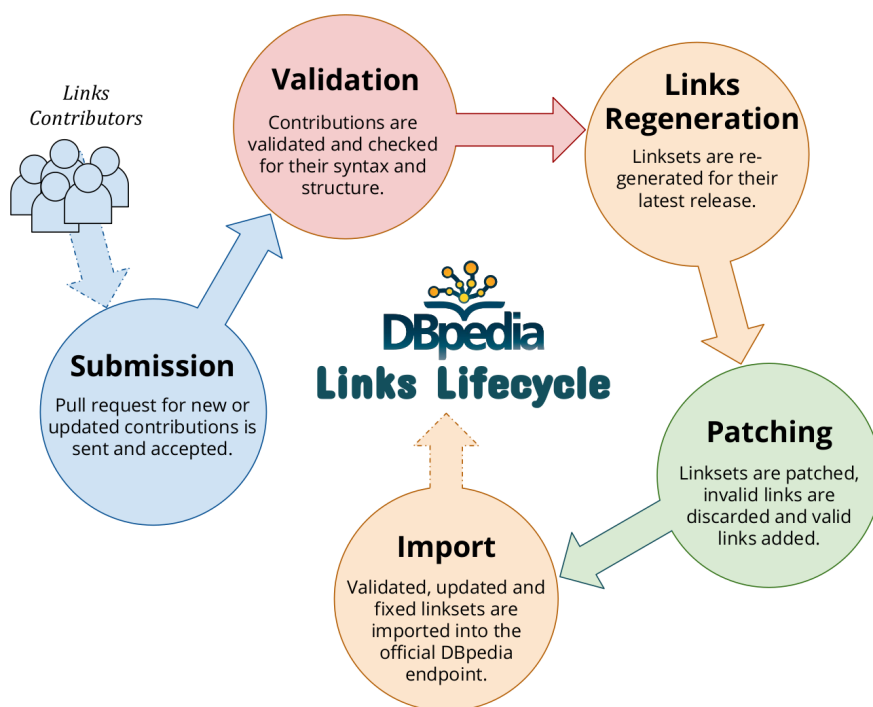


Figura 3.11: Diagrama del ciclo de vida del procesamiento de enlaces de DBpedia [34]

Para el caso de las fechas se realizó una comparación exacta de los atributos *dc:issue* de DBLP contra *dc:date* de *Biblio FIng*. El atributo *dc:date* requirió un pre-procesamiento para extraer el componente del año ya que existían casos en los que el dato del año se encontraba concatenado al mes (ej 2005-10). Para combinar ambas métricas se tomó el mínimo de los resultados parciales de cada comparación, lo que fuerza que las publicaciones coincidan en su año de publicación pero puedan variar mínimamente en el título. En la figura 3.12 se puede ver una representación gráfica de la regla de comparación.

En este caso se compararon 460 publicaciones de la fuente *Biblio FIng* contra 3.9 millones de publicaciones de DBLP. Ambos servicios (*Silk* y *Virtuoso*) corrieron en una computadora con procesador Intel i7-4702HQ con 8 cores y 8GB de RAM. El proceso demoró aproximadamente 4hs y obtuvo 206 enlaces a 194 publicaciones diferentes de DBLP. En la Tabla 3.9 se muestra un resumen de la cantidad de enlaces detectados a los distintos tipos de conceptos. Como se puede observar el 18% de las personas y de los documentos se pudieron asociar a docentes del InCo y publicaciones de DBLP respectivamente.

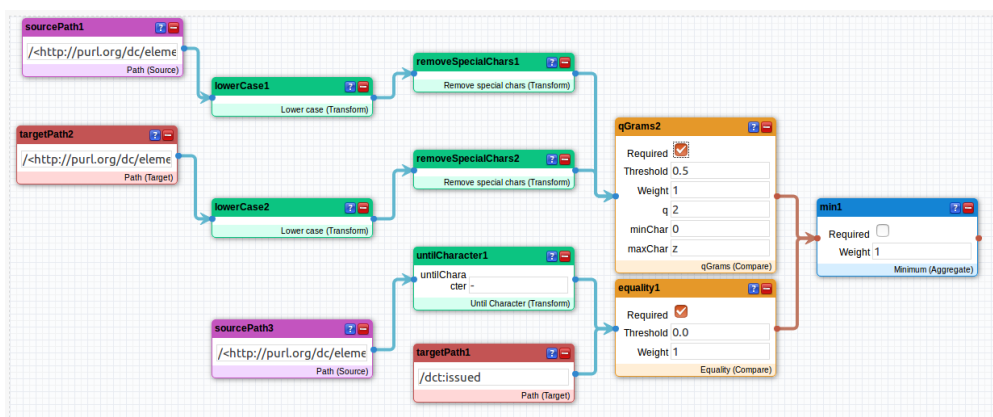


Figura 3.12: Regla de comparación definida en *Silk Workbench*<sup>94</sup> entre publicaciones de *Biblio FIng* y DBLP

## Docentes InCo a autores de DBLP

En este caso las pruebas exploratorias iniciales usando la métrica de n-grama utilizada anteriormente no fueron buenas, ya que se observaron una gran cantidad de falsos positivos. Por tal motivo se optó por utilizar un enfoque iterativo incremental. En este caso se buscó aprovechar el hecho de que el trabajo de investigación es una tarea colaborativa que se manifiesta principalmente por las coautorías de los trabajos publicados. En general los investigadores se organizan en grupos de investigación y muchas de esas coautorías se dan dentro del mismo grupo, así como entre grupos de la misma institución. Esto nos da la pauta de que es bastante probable encontrar docentes de la misma institución en el conjunto de coautores de los docentes previamente identificados. De esta forma se ejecutó un proceso inicial de detección de enlaces utilizando una métrica de comparación muy restrictiva, creando enlaces entre los autores que tengan una coincidencia exacta en el nombre. Una vez que se tuvo un conjunto inicial de (potenciales) autores identificados en DBLP se ejecutó otro proceso de detección de enlaces utilizando una métrica más laxa, pero esta vez entre los docentes del InCo y el conjunto de coautores del grupo de autores identificado inicialmente. En cada iteración se puede ejecutar un proceso de resolución de identidad para detectar posibles casos de ambigüedad que puedan surgir. Esto nos permite utilizar un algoritmo más flexible que aumente las posibilidades de matching pero no genere tantos falsos positivos, gracias a que el conjunto total contra el que se compara es mucho menor. En la Figura 3.13 se puede ver una representación gráfica de la regla de detección.

El proceso de detección de enlaces arrojó 118 enlaces entre los cuales se presen-

Tabla 3.9: Población de las distintas entidades de la fuente *Biblio FIng* asociada a la cantidad enlaces detectados desde la fuente *Docentes InCo* y hacia la fuente DBLP.

Concepto Biblio FING	#Total	#INCO	#DBLP
foaf:Person	492	<sup>a</sup> 91 (18 %)	260 (52 %)
foaf:Document	473	324 (68 %)	<sup>b</sup> 206 (18 %)
bibtex:Conference	180	163 ( 91 %)	123 ( 68 %)
bibtex:Article	137	85 ( 62 %)	58 ( 42 %)
bibtex:Masterthesis	73	19 ( 26 %)	8 ( 11 %)
bibtex:Proceedings	27	20 ( 74 %)	5 ( 19 %)
bibtex:InBook	21	14 ( 67 %)	8 ( 38 %)
bibtex:TechnicalReport	18	18 (100 %)	1 ( 6 %)
bibtex:Book	12	4 ( 33 %)	2 ( 17 %)
bibtex:Publication	4	1 ( 25 %)	1 ( 25 %)

<sup>a</sup> 91 links desde 76 docentes INCO diferentes

<sup>b</sup> 206 links hacia 194 documentos de DBLP diferentes

taron 29 casos de ambigüedad que debieron ser analizados manualmente. Luego del proceso de resolución de identidad quedaron 107 enlaces a diferentes autores de DBLP desde 99 docentes diferentes del InCo. Como se puede ver en la Tabla 3.10 el 63 % del total de docentes del InCo fueron identificados en DBLP. La Figura 3.14 muestra la distribución de docentes por grado discriminando cuántos fueron identificados en DBLP y cuántos no.

## Resolución de identidad

La resolución de identidad es la tarea de desambiguar manifestaciones de entidades del mundo real en distintos registros o menciones. Está relacionada con varias disciplinas diferentes, tales como las bases de datos, la recuperación de información, el procesamiento de lenguaje natural, y más recientemente, el aprendizaje automático, entre otras. La dificultad de esta tarea es variable y puede llegar a requerir desde métodos determinísticos simples hasta métodos probabilísticos más complejos. Muchas veces es imposible realizar la desambiguación de forma automática y se debe recurrir a la intervención humana, por ejemplo utilizando técnicas de *crowdsourcing*. Un claro ejemplo de esto es la plataforma *ResearchGate*, donde resuelven los casos complejos consultando directamente a los usuarios si participan o no en la autoría de una determinada publicación.

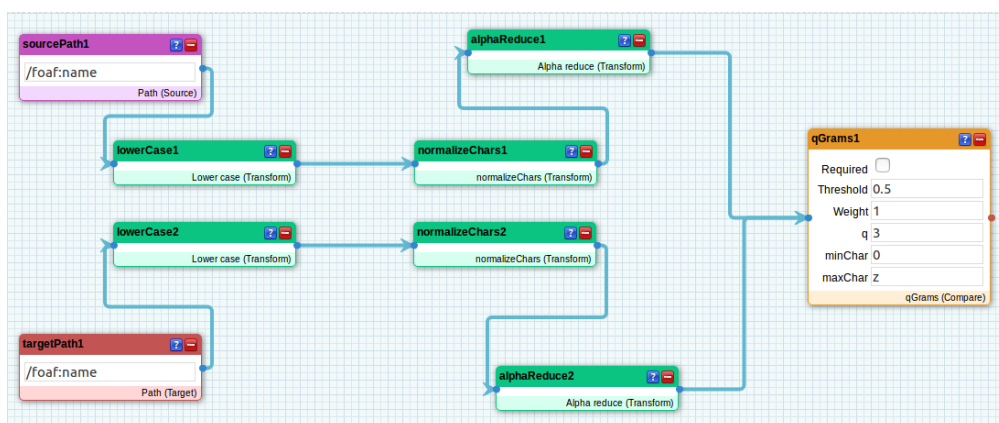


Figura 3.13: Diagrama de regla de comparación entre docentes del InCo y coautores de DBLP

Tabla 3.10: Cantidad de docentes del InCo identificados en DBLP distribuidos por grado.

grado	cantidad de docentes	cantidad de enlaces	%
5	14	14	100 %
4	14	14	100 %
3	49	33	67 %
2	50	26	52 %
1	28	12	42 %
s/d	3	0	0 %
Total:	158	99	63 %

En el caso de los docentes el único atributo que tenemos para comparar es el nombre de la persona. Esto hace que la detección de enlaces sea muy propensa a generar falsos positivos, sobretodo cuando comparamos fuentes de datos tan grandes como DBLP con más de dos millones de personas. Esto hace que utilizar únicamente técnicas de resolución de identidad basadas en atributos no sea suficiente. Una posibilidad es aprovechar la estructura del grafo, es decir, las relaciones representadas en la estructura de los datos. En este sentido un enfoque interesante es el denominado *collective entity resolution*[35]. Esta técnica combina la similitud de atributos con la evidencia relacional y muestra mejores resultados en comparación con enfoques tradicionales. En esta técnica la resolución de identidad se plantea como un problema de agrupamiento (o *clustering*) relacional donde el objetivo es agrupar las referencias de manera que solo aquellas que corresponden a la misma entidad se asignen al mismo grupo. Este tipo de técnicas sacan ventaja de la estructura

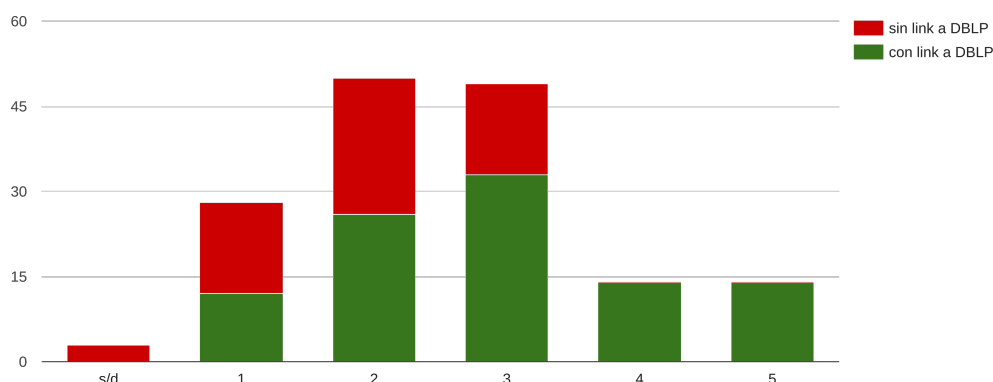


Figura 3.14: Cantidad de docentes del InCo identificados en DBLP distribuidos por grado.

de los datos y permiten mejorar la calidad del proceso de resolución de identidad en las sucesivas iteraciones. Por ejemplo si dos autores tienen nombres similares y una gran cantidad de coautores en común (o coautores que fueron previamente detectados como duplicados) es probable que se trate de la misma persona. Si combinamos esta técnica con la presunción de que es probable que se den coautorías entre docentes de la misma institución, podemos utilizarlo como insumo para resolver casos de falsos positivos. Por ejemplo, en un caso de ambigüedad donde dos autores de DBLP se enlazan al mismo docente InCo por tener nombres muy similares, si el conjunto de coautores difiere más allá de un determinado umbral, los autores estarían ubicados en grupos (o *clusters*) diferentes. En este caso es más probable que el que tenga un índice más alto de coautores dentro del InCo sea el correcto. Si ambos tienen un índice elevado de coautorías es probable que se trate de dos representaciones de la misma persona.

La Tabla 3.11 muestra diversos ejemplos de ambigüedades detectadas luego de la generación de enlaces. Un ejemplo claro es el de la docente *Laura González* quien está registrada en DBLP con el nombre *Laura Gonzales 0001*. Esto provocó que la detección inicial por igualdad de nombre generase un falso positivo al detectar a otra autora homónima registrada con el nombre correcto (*dblp:Laura\_Gonzalez*). En la siguiente iteración con una comparación de nombres más flexible se detectó el enlace a la autora correcta (*dblp:Laura\_González\_0001*) ya que compartía coautorías con otros docentes previamente identificados. Esto generó una ambigüedad pues se tienen dos entidades de DBLP asociadas a la misma entidad de docentes del InCo (*docente:lauragon*). Si analizamos la estructura del grafo de ambas candidatas notamos que la autora correcta tiene 23 publicaciones con 9 coautores detectados dentro del InCo, mientras que el falso positivo no tiene ninguno. A esto se le suma el hecho de que 22 de las 23 publicaciones de DBLP coinciden

Tabla 3.11: Ejemplos de ambigüedades detectadas durante la generación de enlaces entre docentes del InCo y autores de DBLP

Grado	Nombre InCo	Nombre DBLP	#Pub DBLP	#Co-autores <sup>a</sup>	#Pub Biblio <sup>b</sup>
3	Aiala Rosa	Aiala Rosá	6	5 / 11	5
3	Aiala Rosa	Aiala Rosa	1	2 / 3	1
4	Alejandro Gutierrez	Alejandro Gutiérrez	5	5 / 17	0
4	Alejandro Gutierrez	Alejandro Gutierrez	6	1 / 18	0
3	Antonio Lopez Arredondo	Antonio López Arredondo	2	2 / 19	0
3	Antonio Lopez Arredondo	Antonio López	21	1 / 37	0
3	Ariel Sabiguero Yawelak	Ariel Sabiguero	4	3 / 9	0
3	Ariel Sabiguero Yawelak	Ariel Sabiguero Yawelak	2	0 / 0	0
2	Ernesto Dufrechou	Ernesto Dufrechou	16	4 / 19	0
2	Ernesto Dufrechou	Ernesto Dufrechu	2	2 / 6	1
3	Federico Rodriguez	Federico Rodríguez	2	3 / 9	1
3	Federico Rodriguez	Federico Rodríguez-Teja	2	1 / 3	2
2	Fernando Fernandez	Antonio Fernández 0005	1	2 / 7	0
2	Fernando Fernandez	Fernando Fernández	57	1 / 72	0
2	Fernando Fernandez	Fernando Fernandez	5	1 / 9	0
5	Franco Robledo	Franco Robledo	39	7 / 25	3
5	Franco Robledo	Franco Robledo Amoza	15	7 / 15	0
2	Gabriel Lopez	Gabriel López	19	0 / 14	0
2	Gabriel Lopez	Gabriel Lopez	1	0 / 2	0
1	Guillermo Dufort	Guillermo Dufort	1	2 / 8	1
1	Guillermo Dufort	Guillermo Durán	68	1 / 64	0
3	Laura Gonzalez	Laura González 0001	23	9 / 26	22
3	Laura Gonzalez	Laura Gonzalez	1	0 / 4	0
3	Martin Gonzalez	Martin González	6	0 / 9	0
3	Martin Gonzalez	Martín González	1	0 / 4	0
3	Monica Martinez	Monica Martínez	3	1 / 8	0
3	Monica Martinez	Mónica Martínez	2	1 / 6	0
2	Rodrigo Martinez	Rodrigo Martínez	4	2 / 14	0
2	Rodrigo Martinez	Rodrigo Martínez	2	0 / 4	0

<sup>a</sup> Relación de co-autores DBLP identificados como docentes del InCo sobre cantidad total de co-autores DBLP

<sup>b</sup> Cantidad de publicaciones de DBLP identificadas como publicaciones de Biblio FIng

con publicaciones de la fuente *Biblio FIng*. En la Tabla 3.11 también se pueden ver ejemplos con alta probabilidad de ser autores duplicados, tales como *Aiala Rosá* o *Franco Robledo* en donde para cada caso ambos candidatos tienen varios coautores dentro del InCo. En estos casos ambos autores son correctos y a pesar de tener distintos conjuntos de publicaciones representan el mismo autor InCo, en otras palabras son dos entidades de DBLP que representan la misma entidad del mundo real.

### 3.2.4. Publicación

Como se mencionó anteriormente el almacenamiento y gestión de los datos en RDF se hizo utilizando el servidor *Open Link Virtuoso*. Esto nos dio la posibilidad de utilizar el resto de las funcionalidades de dereferenciación y negociación de contenido, así como una terminal SPARQL para realizar consultas. Para esto se establecieron una serie de redirecciones que toman en cuenta la URI base del conjunto de datos y el parámetro **Accept** del encabezado de la solicitud **HTTP** para redirigir a distintos componentes. En la Tabla 3.12 se puede ver un ejemplo

Tabla 3.12: Ejemplo de redirección para la dereferenciación de URIs.

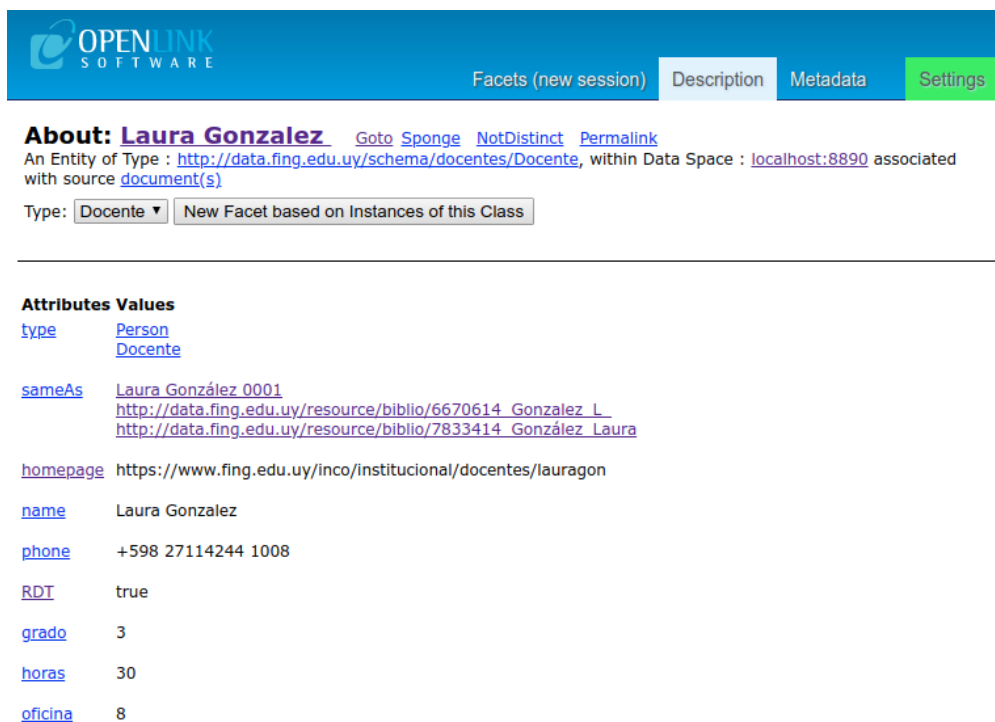
URI	<code>http://data.fing.edu.uy/resource/docente/[userName]</code>
Accept	<code>(text/rdf+n3) (application/rdf+xml) (text/n3) (application/json)</code>
Redirección	<code>/sparql?query=DESCRIBE+&lt;http://data.fing.edu.uy/resource/docente/[userName]&gt;+FROM+&lt;http://data.fing.edu.uy/docentes#&gt;&amp;format=[formato]</code>
URI	<code>http://data.fing.edu.uy/resource/docente/[userName]</code>
Accept	<code>text/html</code>
Redirección	<code>/fct/rdfdesc/description.vsp?g=&lt;http://data.fing.edu.uy/resource/docente/[userName]&gt;&amp;graph=&lt;http://data.fing.edu.uy/docentes#&gt;</code>

de redirección. En caso de que el cliente solicite un formato específico, como por ejemplo `application/rdf+xml`, la solicitud será redirigida a la terminal SPARQL ejecutando el comando `DESCRIBE` para la URI solicitada y el formato indicado en el parámetro `Accept`. Si la solicitud no especifica un formato RDF determinado, se entiende que esta siendo accedida por un ser humano desde un navegador. Esto genera la redirección a otro componente del sistema denominado *FacetBrowser* que muestra los datos de forma amigable en formato HTML. Este componente permite además hacer búsquedas de texto libre o navegar a través de los enlaces y continuar viendo recursos relacionados siguiendo el patrón *follow your nose*<sup>95</sup>. En la Figura 3.15 se puede ver un ejemplo de los datos de un docente accedido desde un navegador y en el Ejemplo 3.3 para el mismo docente se solicitaron los datos en RDF utilizando en el encabezado `Accept` de la solicitud HTTP<sup>96</sup>.

<sup>95</sup><http://patterns.dataincubator.org/book/follow-your-nose.html>

<sup>96</sup>Ejemplo: `curl -H 'Accept: text/n3' http://data.fing.edu.uy/docentes/lauragon`





**About: Laura Gonzalez** [Goto](#) [Sponge](#) [NotDistinct](#) [Permalink](#)

An Entity of Type : <http://data.fing.edu.uy/schema/docentes/Docente>, within Data Space : <localhost:8890> associated with source [document\(s\)](#)

Type:

---

**Attributes Values**

<a href="#">type</a>	<a href="#">Person</a> <a href="#">Docente</a>
<a href="#">sameAs</a>	<a href="#">Laura González 0001</a> <a href="http://data.fing.edu.uy/resource/biblio/6670614_Gonzalez_L">http://data.fing.edu.uy/resource/biblio/6670614_Gonzalez_L</a> <a href="http://data.fing.edu.uy/resource/biblio/7833414_González_Laura">http://data.fing.edu.uy/resource/biblio/7833414_González_Laura</a>
<a href="#">homepage</a>	<a href="https://www.fing.edu.uy/inco/institucional/docentes/lauragon">https://www.fing.edu.uy/inco/institucional/docentes/lauragon</a>
<a href="#">name</a>	Laura Gonzalez
<a href="#">phone</a>	+598 27114244 1008
<a href="#">RDT</a>	true
<a href="#">grado</a>	3
<a href="#">horas</a>	30
<a href="#">oficina</a>	8

Figura 3.15: Ejemplo de visualización de los datos de un docente accediendo desde un navegador.

Ejemplo 3.3: Ejemplo de visualización de los datos de un docente en RDF serializado en formato N3.

```

1  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
2  @prefix owl: <http://www.w3.org/2002/07/owl#> .
3  @prefix ns1: <http://data.fing.edu.uy/resource/docente/> .
4  @prefix foaf: <http://xmlns.com/foaf/0.1/> .
5  ns1:lauragon rdf:type foaf:Person .
6  @prefix ns2: <http://data.fing.edu.uy/resource/biblio/>
7  @prefix ns3: <http://data.fing.edu.uy/schema/docentes/> .
8  @prefix ns4: <http://dblp.13s.de/d2r/resource/authors/>
9  ns1:lauragon rdf:type ns3:Docente ;
10  foaf:homepage "https://www.fing.edu.uy/inco/institucional/docentes/
11  lauragon" ;
12  foaf:name "Laura Gonzalez" ;
13  foaf:phone "+598 27114244 1008" ;
14  ns3:RDT "true" ;
15  ns3:grado "3" ;
16  ns3:horas "30" ;
17  ns3:oficina "8" ;
18  owl:sameAs ns4:Laura_González_0001 ;

```

```
18 owl:sameAs ns2:6670614_Gonzalez_L_ ;
19 owl:sameAs ns2:7833414_González_Laura .
```

Una vez que se tienen resueltos los mecanismos de consulta y dereferenciación es momento de anunciar los datos. Con el fin de facilitar y promover el uso de los datos es recomendable agregar metadatos que describan el conjunto de datos. Estos metadatos pueden incluir información de autoría, frecuencia de actualización, la licencia de los datos o inclusive enlaces a otros recursos como ejemplos u ontologías. En este sentido uno de los principales mecanismos es la inclusión de una descripción del conjunto de datos utilizando el vocabulario VoID<sup>97</sup>. En este caso se crearon descripciones VoID con metadatos básicos para los conjuntos de datos generados.

### 3.2.5. Análisis de resultados

Para mostrar el potencial uso de los datos se realizó un análisis básico que explota los principios de datos enlazados. Para esto se realizaron algunas consultas SPARQL que aprovechan la visión integrada de los datos utilizando los enlaces detectados entre las distintas fuentes.

Para la creación y ejecución de las consultas se utilizó el cliente SPARQL *YASGUI*<sup>98</sup>. Esta aplicación web se conecta a un terminal SPARQL y proporciona un área de texto con funcionalidades tales como, resaltado de sintaxis y autocompletado que facilitan la creación de las consultas. *YASGUI* proporciona además una interfaz amigable para la creación de diversos tipos de gráficas a partir de los resultados de las consultas. En el Apéndice B se pueden encontrar algunos ejemplos de las consultas SPARQL creadas para analizar los datos.

### Cantidad de publicaciones y coautorías de los docentes

En las Figuras 3.16 y 3.17 se muestra la distribución de publicaciones y coautorías en función del grado del docente y el tipo de publicación. Allí se puede observar una relación bastante clara entre el grado del docente y la cantidad de publicaciones y coautorías dentro de la institución. Esto podría explicarse ya que un docente de mayor grado suele tener una mayor trayectoria y por ende una mayor acumulación

<sup>97</sup><https://www.w3.org/TR/void/>

<sup>98</sup><http://about.yasgui.org/>

	grado	1	2	3	4	5	Totals
tipo							
Article		1.50	2.36	3.84	11.79	16.00	7.58
InCollection				1.00	4.67	1.00	2.57
InProceedings		1.45	3.79	5.77	18.57	17.57	8.43
PhDThesis				1.00	1.00	1.00	1.00
Totals		1.47	3.26	4.67	13.75	14.77	7.61

Figura 3.16: Promedio de publicaciones por docente y tipo de publicación.

	grado	1	2	3	4	5	Totals
tipo							
Article		1.25	1.21	2.05	2.64	3.23	2.19
InCollection				0.00	0.33	1.00	0.57
InProceedings		1.91	2.42	2.90	5.21	5.86	3.45
PhDThesis				0.00	0.00	0.00	0.00
Totals		1.73	1.97	2.35	3.47	4.10	2.74

Figura 3.17: Promedio de coautorías dentro del InCo por docente y tipo de publicación.

de publicaciones. A su vez posibilita que trabaje en colaboración con una mayor cantidad y diversidad de coautores a lo largo de su vida. Esto nos da la pauta que es más eficiente priorizar los docentes de mayor grado en las etapas tempranas de detección de enlaces o resolución de identidad, ya que nos proporcionan mayor información para las iteraciones posteriores.

	cantidad	0	5	10	15	20	25	30	35	45	50	55	60	70	75	Totals
grado																
1		11	1													12
2		19	4	1	2											26
3		18	8	1	2	3								1		33
4		3	1	3	1	2	1	1	1					1		13
5		2	2	1	3	1	1	1	1	1	1			1	1	14
Totals		50	16	5	8	4	5	2	1	1	1	1	2	1	1	98

Figura 3.18: Histograma de cantidad de publicaciones por grado en intervalos de 5 publicaciones.

	grado	1	2	3	4	5	Totals
tipo							
Article		0.58	2.13	5.12	17.05	16.18	12.23
InCollection				0.00	6.35	0.00	4.16
InProceedings		0.69	3.40	7.55	15.10	10.87	10.56
PhDThesis				0.00	0.00	0.00	0.00
Totals		0.64	3.05	6.53	15.67	13.62	10.95

Figura 3.19: Desviación estándar del promedio de coautorías entre docentes del InCo por grado y tipo de publicación.

A su vez esta información se puede usar de forma complementaria para buscar

valores atípicos en la etapa de resolución de identidad o de control de calidad. Por ejemplo, evaluando que tanto se alejan los candidatos de los valores medios para el grado al que pertenecería. Se puede ver un ejemplo de esto en la ambigüedad detectada para el docente *Guillermo Dufort* de la Tabla 3.11, donde el falso positivo tiene 68 publicaciones, lo que se aleja mucho del promedio para un docente grado 1. Por otro lado las Figuras 3.18 y 3.19 muestran la dispersión de cantidad de coautorías y publicaciones según el grado del docente. Esto nos indica que se debe tener cuidado ya que, además de aumentar, los valores tienden a estar más dispersos conforme aumenta el grado del docente. Lo anterior nos indica que estas métricas suelen ser más confiables en docentes de grado bajo.

De igual forma en el proceso de control de calidad el docente *Eduardo Fernández* llamó la atención por la gran cantidad de publicaciones asociadas (133). Estos datos atípicos se contrastaron manualmente contra otras fuentes (como por ejemplo CVuy<sup>99</sup> y el sitio web de DBLP) y se logró detectar que se debían a errores en la fuente de datos que agrupaba las publicaciones de varias personas. Si bien en la versión oficial de DBLP<sup>100</sup> figuran como personas distintas<sup>101 102 103</sup>, en la versión DBLP++ publicada como Datos Enlazados estas personas fueron unificadas bajo la misma URI `dblp:Eduardo_Fern%C3%A1ndez`. Esto hace que todas sus publicaciones se unifiquen y distorsionen los datos. Este error fue solucionado en la versión de octubre de 2018 que publica DBLP++.

## Datos generales sobre los tipos de publicaciones encontradas

En lo concerniente a los tipos de publicaciones en la Figura 3.20 se puede ver que la mayoría de las publicaciones se realizan en congresos (56,7%), seguidas de cerca por los artículos en revistas (40.9%). La misma tendencia se puede ver si las distribuimos por grado de docente o por año (Figura 3.21).

## Series temporales

Consultando la propiedad `dct:issue` de las publicaciones tenemos acceso al año en que se publicó cada trabajo. Esto nos permite hacer estudios de series temporales y

---

<sup>99</sup><https://buscadorcvuy.anii.org.uy>

<sup>100</sup><http://dblp.org>

<sup>101</sup>[https://dblp.org/pers/hd/f/Fern=acute=ndez\\_0001:Eduardo](https://dblp.org/pers/hd/f/Fern=acute=ndez_0001:Eduardo)

<sup>102</sup>[https://dblp.org/pers/hd/f/Fern=acute=ndez\\_0002:Eduardo](https://dblp.org/pers/hd/f/Fern=acute=ndez_0002:Eduardo)

<sup>103</sup>[https://dblp.org/pers/hd/f/Fern=acute=ndez\\_0003:Eduardo](https://dblp.org/pers/hd/f/Fern=acute=ndez_0003:Eduardo)

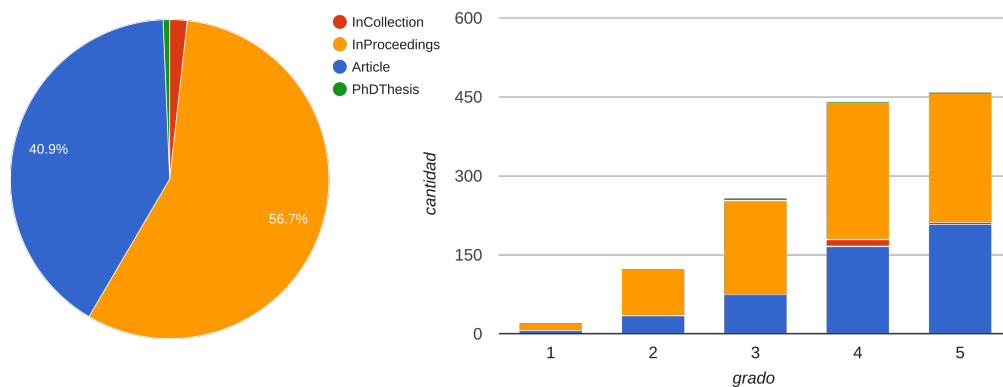


Figura 3.20: Distribución de publicaciones por tipo (izquierda) y por tipo y grado (derecha)

evaluar la evolución a lo largo del tiempo de distintos aspectos. En la Figura 3.21 se muestra la evolución en la cantidad de publicaciones por tipo a lo largo del tiempo. Como se puede observar en la Figura 3.22 dicha distribución se ajusta bastante a la misma distribución pero aplicada a todas las publicaciones de DBLP.

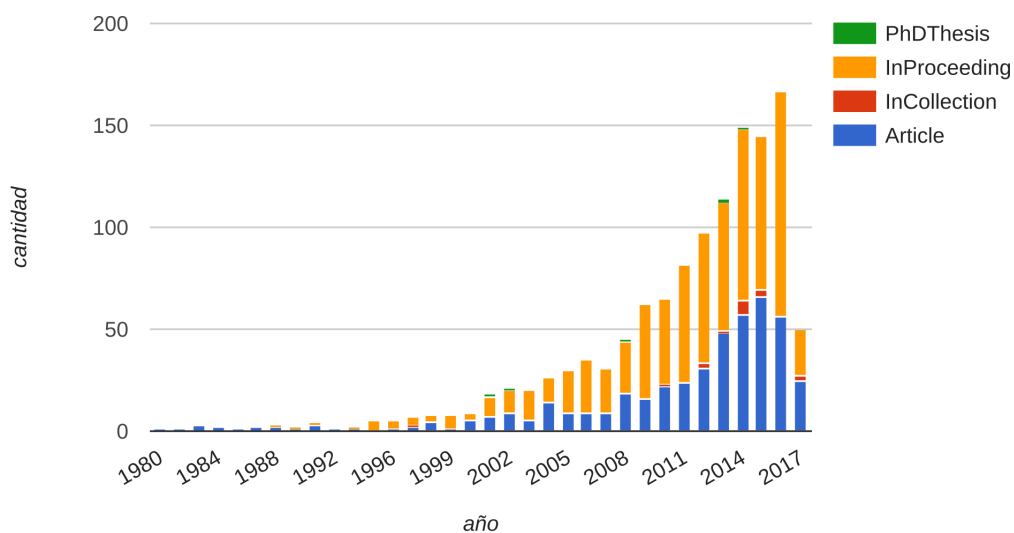


Figura 3.21: Distribución publicaciones del InCo por año y tipo

### Lugares de publicación más populares

El análisis de los lugares de publicación también es interesante ya que nos da una pauta de las comunidades científicas más populares en las que se participa.

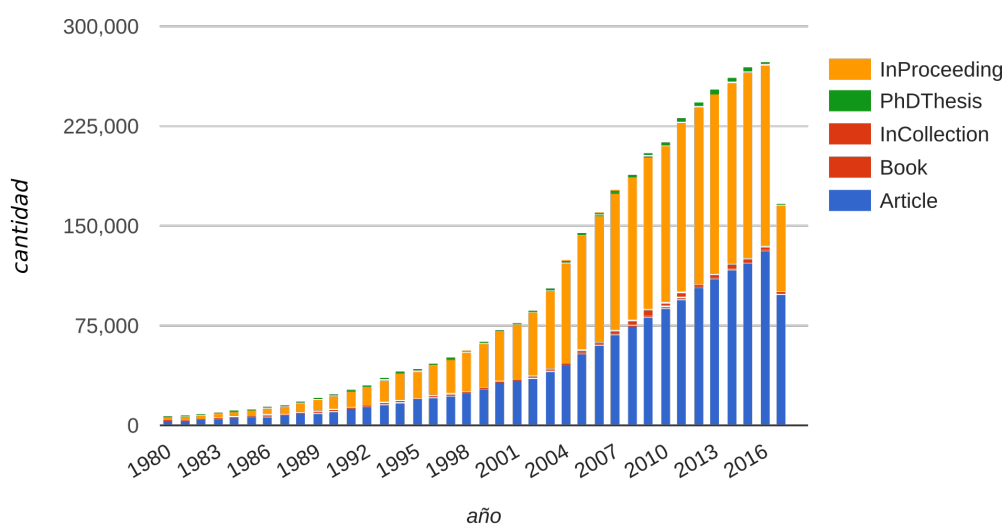


Figura 3.22: Distribución de todas las publicaciones de DBLP por año y tipo

Como se puede ver en la Tabla 3.13 claramente en el primer puesto se encuentra las conferencias y la revista del CLEI. Esto tiene sentido dado que es un evento regional que engloba muchas temáticas diferentes, lo cual facilita el acceso y la participación de los investigadores locales.

### Limitaciones

Si bien el objetivo del trabajo no es realizar un análisis estadístico riguroso, es importante tener en cuenta que existen varias limitaciones con respecto a los datos que pueden afectar los resultados.

**Falsos Positivos.** Existe la posibilidad de introducir datos erróneos al tomar en cuenta las publicaciones de autores que fueron falsos positivos en la etapa de generación de enlaces. Este problema se puede mitigar mejorando los controles en la etapa de resolución de identidad.

**Datos incorrectos en las fuentes.** Como se mencionó anteriormente las fuentes de datos obtienen gran parte de sus datos procesando y analizando sitios web mediante el uso de técnicas de *Information Retrieval*. Este tipo de técnicas muchas veces se basan en el procesamiento de datos semi-estructurados o no estructurados, lo que puede generar errores. Un ejemplo de esto son los autores duplicados detectados en la fuente DBLP. Este tipo de errores se puede atacar integrando otras fuentes que agreguen redundancia de datos y den la posibilidad de desarrollar

Tabla 3.13: Top 20 de conferencias y revistas con más publicaciones en el período 2006-2016

Conferencias			Revistas	
Ranking	Nombre	Totales	Nombre	Totales
1	CLEI	27	CLEI Electron. J.	23
2	IWINAC	22	CoRR	19
3	CARLA	14	Neurocomputing	17
4	LANOMS	9	IRE Trans. Information Theory	11
5	SCCC	9	IJMHeur	8
6	ISIT	9	European Journal of Operational Research	8
7	LANC	7	Electr. Notes Theor. Comput. Sci.	8
8	ISW-LOD@IBERAMIA	6	ITOR	7
9	ICUMT	5	Int. J. Neural Syst.	6
10	MEDINFO	5	Sci. Comput. Program.	6
11	ICIP	5	Electronic Notes in Discrete Mathematics	6
12	CIAPR	5	Computers & OR	5
13	ReMo2V	5	The Journal of Supercomputing	5
14	JISIC	5	Computers & Graphics	5
15	OTM Workshops	5	Cluster Computing	4
16	SBLP	5	International Journal of Man-Machine Studies	3
17	3PGCIC	5	Algorithmica	3
18	IPPS/SPDP	4	Annals OR	3
19	SCC	4	Appl. Soft Comput.	3
20	SBMF	4	SIAM J. Imaging Sciences	3

procesos de evaluación de calidad más completos.

**Datos parciales.** Como se muestra en la Tabla 3.14 se logró identificar en DBLP al 63% de los docentes del InCo, lo que implica que el 37% restante no se está tomando en cuenta en las consultas. Por otro lado los docentes identificados se concentran en los grados más altos que suelen tener una mayor cantidad de publicaciones. Esto nos da la pauta de que por más que el porcentaje de docentes identificados no sea muy alto, estos concentran un porcentaje elevado de las publicaciones del instituto.

**Datos parciales en series temporales.** Una limitación importante es la falta de información respecto a los períodos de actividad de los docentes del InCo, tanto para ex-docentes como para docentes activos. Esto es particularmente importante para el análisis de series temporales. El hecho de no contar con el histórico de docentes hace que no se tomen en cuenta ninguno de sus trabajos. Lo mismo ocurre a la inversa cuando un docente se incorpora a la plantilla, todas las publicaciones que pueda haber publicado con otras instituciones serán tomadas en cuenta. Por ejemplo el ex-docente *Daniel Perovich* tiene publicaciones realizadas durante su trabajo en el InCo, pero no serán tomadas en cuenta ya que actualmente trabaja en la Universidad de Chile y no forma parte de la lista de docentes. Para evitar esto se deben tener los períodos de actividad en la institución para cada docente así como la planilla histórica. Esto permitiría ajustar las consultas para tomar en cuenta únicamente las publicaciones publicadas en su período de actividad.

# Capítulo 4

## Conclusiones y trabajos futuros

### 4.1. Conclusiones

Como parte del presente trabajo se realizó una revisión de las metodologías, recomendaciones y buenas prácticas existentes para la publicación de Datos Enlazados en la web. Estas guías y recomendaciones se utilizaron como base para el análisis de dos casos de estudio de características diferentes, como lo son CNX.org y producción bibliográfica del InCo. Al igual que otros procesos de análisis y gestión de información, la publicación de Datos Enlazados requiere un esfuerzo considerable. Por tal motivo es natural que el equipo de trabajo ponga un mayor énfasis en los aspectos relevantes a su área de conocimiento o a sus requerimientos particulares. Este proceso incremental se enmarca en el denominado ciclo de vida de los Datos Enlazados, que se puede ver como una iteración sobre las distintas etapas, con el fin de ir enriqueciendo paulatinamente la calidad y la semántica del conjunto de datos. El esfuerzo invertido en cada una de las actividades durante la publicación va a depender fundamentalmente de la naturaleza de los datos, la disponibilidad y madurez de las herramientas y la experiencia del equipo de trabajo. Esto se vio reflejado en los dos casos de estudio, en donde se requirió diferente énfasis en las etapas ejecutadas.

En el caso de CNX.org se hizo mayor hincapié en la etapa de especificación por tratarse de una plataforma desconocida que proporciona varios formatos y mecanismos de acceso a los datos. Lamentablemente no se identificaron autores del InCo, sin embargo, el hecho de usar el protocolo OAI-PMH nos da la posibilidad de reutilizar el mismo mecanismo de publicación en otros sistemas que provean



dicho protocolo, como por ejemplo DSpace<sup>104</sup>, que es muy utilizado en repositorios académicos.

En el segundo caso se publicaron como Datos Enlazados la lista de docentes y la base bibliográfica disponibles en el sitio web de FIng. Con el objetivo de integrarlas a la fuente de información bibliográfica en ciencias de la computación DBLP, se diseñaron y ejecutaron procesos de detección de enlaces y resolución de identidad entre las tres fuentes: *Docentes del InCo*, entidades de *Biblio FIng*, y entidades de DBLP, tal como se muestra en la Figura 3.10. Este proceso expuso varios de los desafíos relacionados con la integración de datos como por ejemplo la resolución de integridad entre entidades con pocos atributos para comparar. Contar con la base de publicaciones ampliada y distribuida, nos dio la posibilidad de ejecutar consultas complejas, que explotan las características estructurales de los Datos Enlazados. De esta forma se pudo obtener y analizar una gran cantidad de información dispersa, relacionada con la producción científica de los docentes del InCo.

#### 4.1.1. Trabajos similares

*Publishing Bibliographic Data on the SemanticWeb using BibBase*[36]. *BibBase*<sup>105</sup> es un sistema orientado a investigadores pensado para gestionar información bibliográfica. Proporcionando el identificador correspondiente, permite importar los datos de fuentes existentes, tales como DBLP, *BibSonomy*<sup>106</sup> o *Mendeleys*<sup>107</sup>, o cargarlos a partir de archivos BibTex. Dicho trabajo también maneja datos bibliográficos representados de diferente manera, y provenientes de diversas fuentes. A su vez, enfrenta problemas de integración de datos y resolución de identidad, similares a los abordados en esta tesis.

*Linked Data for Learning Analytics: The Case of the LAK Dataset*[37]. El trabajo explora las oportunidades que presenta el uso de Datos Enlazados en el campo de *Learning Analytics*. Para esto analiza un conjunto de publicaciones presentados en las conferencias relacionadas con *Learning Analytics* y *Educational Datamining*, y explota los enlaces a otros juegos de datos para obtener nuevo conocimiento. El caso de estudio de producción bibliográfica del InCo, en particular el análisis de los datos, está fuertemente inspirado en este artículo.

Si bien el presente trabajo toma varias ideas presentadas en los dos trabajos men-

---

<sup>104</sup><https://duraspace.org/dspace/>

<sup>105</sup><https://bibbase.org/>

<sup>106</sup><https://www.bibsonomy.org/>

<sup>107</sup><https://www.mendeley.com/>

cionados, se diferencia particularmente en que hace foco en el estudio del proceso general y las etapas involucradas en el ciclo de vida de los Datos Enlazados.

## 4.2. Trabajo Futuro

El presente estudio es apenas una prueba de concepto de que es posible exponer e integrar diversas fuentes de datos utilizando los principios de los Datos Enlazados. Al día de hoy existen varias metodologías que sirven como guía para la publicación, así como un conjunto importante de herramientas que dan soporte a las distintas etapas. Durante el desarrollo de este proyecto se identificaron varias vetas para trabajos futuros, que se abren en torno al área investigada y los casos de estudio analizados. A continuación se listan algunas de esas líneas de trabajo identificadas.

**Fuentes de Datos.** Analizar e integrar otras fuentes de datos nacionales con información académica como por ejemplo:

- CVuy de la **Agencia Nacional de Investigación e Innovación (ANII)**: fuente de información con los perfiles de muchos de los investigadores del país, así como información del área de investigación, las publicaciones y proyectos en los que participó. Permitiría enriquecer enormemente la información del perfil académico de los docentes con datos como: fecha de obtención PhD, Master, o grados docentes; información sobre su grupo, área o proyectos de investigación, etc. Esto agregaría nuevas dimensiones de análisis y la posibilidad de crear nuevos y mejores servicios de recomendación o búsqueda semántica. Actualmente CVuy proporciona acceso vía Webservice a su base de datos pero está restringido al uso interno de las instituciones.
- Proyecto Colibrí<sup>108</sup>: es el repositorio institucional de la producción científica y académica de la Universidad de la República. Su característica de abierto y colaborativo lo perfilan como un repositorio de referencia en el futuro. Está desarrollado utilizando la plataforma DSpace<sup>109</sup>. Una posibilidad es habilitar la interfaz OAI-PMH, soportada por la versión en uso actualmente<sup>110</sup>, y utilizar un enfoque similar al propuesto en el caso de CNX.org. Otra opción es migrar a la versión 5.0 de esta plataforma que provee soporte nativo para Datos Enlazados<sup>111</sup>.

---

<sup>108</sup><https://www.colibri.udelar.edu.uy/>

<sup>109</sup><http://www.dspace.org/>

<sup>110</sup><https://wiki.duraspace.org/display/DSDOC4x/OAI>

<sup>111</sup><https://jira.duraspace.org/browse/DS-2061>

- Obtener e integrar la planilla de la historia laboral de docentes con sus fechas de actividad para poder contar con estadísticas más precisas.

**Clasificación.** Explorar el uso de técnicas de extracción de entidades (*Named Entity Recognition*) para analizar el texto libre que puedan contener algunas propiedades. Algunos ejemplos de esto pueden ser el reconocimiento de términos de alguna taxonomía de clasificación (como *Library of Congress Subject Headings*) en las palabras clave, o la identificación de entidades de una ontología de propósito general (como DBepedia) en los resúmenes (*abstracts*) de las publicaciones. Contar con información organizada y categorizada bajo una misma taxonomía nos permitiría realizar análisis más ricos explotando las relaciones explícitas entre los conceptos, o detectando relaciones implícitas mediante el uso de inferencia. Por ejemplo, buscar posibles puntos de contacto entre investigadores de disciplinas diferentes que tienen temas de investigación o intereses en común, o descubrir investigaciones que se relacionen a una cierta entidad o temática particular.

**Análisis de los datos.** Aprovechar la integración con otras fuentes para realizar análisis más complejos que generen nuevo conocimiento. Por ejemplo detectar la evolución de los temas de investigación o capacidades de los investigadores locales a lo largo del tiempo. Se podría generar un perfil semántico detallado de cada investigador y poder hacer clústers de afinidad entre investigadores de la Universidad.

**Explotar otros usos.** Por ejemplo evaluar la posibilidad de cargar los datos en la plataforma de gestión de datos académicos VIVO<sup>112</sup>. Esta herramienta se basa en las tecnologías de la web semántica y permite almacenar y vincular perfiles detallados de investigadores, publicaciones, proyectos, entre otros. Proporciona un entorno potente y amigable para la búsqueda y visualización de la información de investigación de un institución. La plataforma permite integrar los datos con otras universidades, y dado que respeta mismos principios de los Datos Enlazados, los datos quedan accesibles en la web de datos. La integración con esta plataforma podría implicar la alineación de los vocabularios utilizados en este documento con los que forman parte del modelo de datos de VIVO. Esto tendría aún más impacto si también se integran los datos del SNI<sup>113</sup> de la ANII. De esta forma se podría crear un registro de expertos o investigadores en determinados temas, tener una visión integral y transparente de la investigación y conocimiento científico a nivel nacional.

---

<sup>112</sup><http://vivoweb.org>

<sup>113</sup><http://sni.org.uy>

# Bibliografía

- [1] Tim Berners-Lee. *The original proposal of the WWW*. 1989. URL: <https://www.w3.org/History/1989/proposal.html> (visitado 15-02-2018).
- [2] Tim Berners-Lee. *The World Wide Web: Past, Present and Future*. 1996. URL: <https://www.w3.org/People/Berners-Lee/9610-IEEE-Computer-v1.html> (visitado 15-02-2018).
- [3] J. Hendler y O. Lassila T. Berners-Lee. “The Semantic Web”. En: *Scientific American* 284.May (2001), págs. 34-43.
- [4] Leo Sauermann y Richard Cyganiak. *Semantic Web Case Studies and Use Cases*. 2012. URL: <https://www.w3.org/2001/sw/sweo/public/UseCases/> (visitado 13-02-2018).
- [5] Tim Berners-Lee. *Semantic Web - XML2000*. 2000. URL: <https://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview.html> (visitado 01-11-2018).
- [6] World Wide Web Consortium (W3C). *World Wide Web Consortium Process Document*. 2005. URL: <https://www.w3.org/2018/Process-20180201/> (visitado 20-02-2018).
- [7] Thomas R. Gruber. “Toward principles for the design of ontologies used for knowledge sharing”. En: *International Journal of Human - Computer Studies* 43.5-6 (1995), págs. 907-928. ISSN: 10959300. DOI: [10.1006/ijhc.1995.1081](https://doi.org/10.1006/ijhc.1995.1081). arXiv: [0701907v3 \[math\]](https://arxiv.org/abs/0701907v3).
- [8] Nicola Guarino y col. “Handbook on Ontologies”. En: (2009), págs. 1-17. ISSN: 10744770. DOI: [10.1007/978-3-540-92673-3](https://doi.org/10.1007/978-3-540-92673-3). arXiv: [arXiv:1011.1669v3](https://arxiv.org/abs/1011.1669v3). URL: <http://link.springer.com/10.1007/978-3-540-92673-3>.
- [9] Richard Cyganiak, David Wood y Markus Lanthaler. “RDF 1.1 Concepts and Abstract Syntax”. En: *W3C Recommendation 25 February 2014* (2014). ISSN: 13514180. DOI: [10.1007/s13398-014-0173-7.2](https://doi.org/10.1007/s13398-014-0173-7.2). arXiv: [arXiv:1011.1669v3](https://arxiv.org/abs/1011.1669v3).
- [10] Dan Brickley y R.V. V Guha. *RDF Schema 1.1*. 2014. URL: <http://www.w3.org/TR/rdf-schema/>.

- [11] D. L. McGuinness y F. Van Harmelen. “OWL Web Ontology Language Overview”. En: *W3C recommendation* 10 (2004), págs. 1-22. ISSN: 15302180. DOI: [10.1145/1295289.1295290](https://doi.org/10.1145/1295289.1295290).
- [12] Andy Seaborne Eric Prud’hommeaux. *SPARQL Query Language for RDF*. 2008. URL: <https://www.w3.org/TR/rdf-sparql-query/> (visitado 01-02-2018).
- [13] The W3C SPARQL Working Group. *SPARQL 1.1 Protocol*. 2013. URL: <https://www.w3.org/TR/2013/REC-sparql11-protocol-20130321/> (visitado 17-02-2018).
- [14] Steve Harris y Andy Seaborne. *SPARQL 1.1 Overview*. 2013.
- [15] Christian Bizer, Tom Heath y Tim Berners-Lee. “Linked Data - The Story So Far”. En: *International Journal on Semantic Web and Information Systems* (2009). ISSN: 1552-6283. DOI: [10.4018/jswis.2009081901](https://doi.org/10.4018/jswis.2009081901). arXiv: [1011.1669](https://arxiv.org/abs/1011.1669).
- [16] Alistair Miles y Sean Bechhofer. *SKOS Simple Knowledge Organization System Reference*. 2009. URL: <http://www.w3.org/TR/skos-reference/>.
- [17] Mikael Nilsson. “From Interoperability to Harmonization in Metadata Standardization”. Tesis doct. 2010. URL: <http://www.diva-portal.org/smash/get/diva2:369527/FULLTEXT02>.
- [18] Government Linked Data Working Group y W3C. *Best Practices for Publishing Linked Data*. 2014. URL: <https://www.w3.org/TR/ld-bp/> (visitado 01-11-2018).
- [19] Bernadette Hyland y David Wood. *Linking Government Data*. Ed. por David Wood. New York, NY: Springer New York, 2011, págs. 3-26. ISBN: 978-1-4614-1766-8. DOI: [10.1007/978-1-4614-1767-5](https://doi.org/10.1007/978-1-4614-1767-5). arXiv: [arXiv:1011.1669v3](https://arxiv.org/abs/1011.1669v3). URL: <http://link.springer.com/10.1007/978-1-4614-1767-5>.
- [20] Michael Hausenblas. *Linked data life cycles*. 2011. URL: <https://www.slideshare.net/mediasemanticweb/linked-data-life-cycles>.
- [21] Boris Villazón Terrazas y col. “Methodological Guidelines for Publishing Government Linked Data”. En: *Linking government data*. (2011), págs. 27-49. DOI: [10.1007/978-1-4614-1767-5](https://doi.org/10.1007/978-1-4614-1767-5).
- [22] W3C. *Cool URIs for the Semantic Web*. 2008. URL: <http://www.w3.org/TR/cooluris/> (visitado 10-02-2018).
- [23] Sören Auer y col. “Managing the life-cycle of linked data with the LOD2 stack”. En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7650 LNCS.PART 2 (2012), págs. 1-16. ISSN: 03029743. DOI: [10.1007/978-3-642-35173-0-1](https://doi.org/10.1007/978-3-642-35173-0-1).

- [24] Tom Heath y Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Vol. 1. 2011, págs. 1-136. ISBN: 9781608454303. DOI: [10.2200/S00334ED1V01Y201102WBE001](https://doi.org/10.2200/S00334ED1V01Y201102WBE001).
- [25] Leigh Dodds y col. *Linked data patterns*. 2011, pág. 79. URL: <http://patterns.dataincubator.org/book/linked-data-patterns.pdf><http://patterns.dataincubator.org/book/index.html>.
- [26] Ivan Ermilov, Jens Lehmann y Michael Martin. “LODStats: The Data Web Census Dataset”. En: 9982 (2016), págs. 38-46. ISSN: 03029743. DOI: [10.1007/978-3-319-46547-0](https://doi.org/10.1007/978-3-319-46547-0). URL: <http://link.springer.com/10.1007/978-3-319-46547-0>.
- [27] Pierre Yves Vandenbussche y col. “SPARQLES: Monitoring public SPARQL endpoints”. En: *Semantic Web 8.6* (2017), págs. 1049-1065. ISSN: 22104968. DOI: [10.3233/SW-170254](https://doi.org/10.3233/SW-170254).
- [28] Max Schmachtenberg, Christian Bizer y Heiko Paulheim. “Adoption of the linked data best practices in different topical domains”. En: *The Semantic Web-ISWC ...* (2014). DOI: [10.1007/978-3-319-11964-9\\_16](https://doi.org/10.1007/978-3-319-11964-9_16).
- [29] Rhaptos. *OpenArchivesHarvesting - Rhaptos-Trac*. URL: <https://trac.rhaptos.org/wiki/OpenArchivesHarvesting> (visitado 20-08-2017).
- [30] Bernhard Haslhofer y Bernhard Schandl. “Interweaving OAI-PMH data sources with the linked data cloud”. En: *International Journal of Metadata, Semantics ...* (2010), pág. 15. URL: <http://inderscience.metapress.com/index/t770153631427210.pdf>.
- [31] Bernhard Haslhofer y Bernhard Schandl. *The OAI2LOD Server: Exposing OAI-PMH metadata as linked data*. 2008. URL: <http://www.ra.ethz.ch/cdstore/www2008/events.linkeddata.org/ldow2008/papers/03-haslhofer-schandl-oai2lod-server.pdf>.
- [32] Mikhail Bilenko, Beena Kamath y Raymond J. Mooney. “Adaptive blocking: Learning to scale up record linkage”. En: *Proceedings - IEEE International Conference on Data Mining, ICDM*. 2006. ISBN: 0769527019. DOI: [10.1109/ICDM.2006.13](https://doi.org/10.1109/ICDM.2006.13).
- [33] Gabriel Recchia y Max Louwerse. “A Comparison of String Similarity Measures for Toponym Matching”. En: *Acm Sigspatial c* (2013), págs. 1-8. DOI: [10.1145/2534848.2534850](https://doi.org/10.1145/2534848.2534850).
- [34] Milan Dojchinovski, Robert Rößling y Sebastian Hellmann. “DBpedia Links : The Hub of Links for the Web of Data”. En: (2016), págs. 11-14.
- [35] Indrajit Bhattacharya y Lise Getoor. “Collective entity resolution in relational data”. En: *ACM Transactions on Knowledge Discovery from Data* 1.1 (2007), 5-es. ISSN: 15564681. DOI: [10.1145/1217299.1217304](https://doi.org/10.1145/1217299.1217304). URL: <http://portal.acm.org/citation.cfm?doid=1217299.1217304>.

- [36] Reynold S. Xin y col. “Publishing bibliographic data on the semantic web using bibbase”. En: *CEUR Workshop Proceedings*. 2010. DOI: [10.3233/SW-2012-0062](https://doi.org/10.3233/SW-2012-0062).
- [37] Davide Taibi y Stefan Dietze. “Linked Data for Learning Analytics: The Case of the LAK Dataset”. En: *Handbook of Learning Analytics* (2017), págs. 337-345. DOI: [10.18608/hla17.029](https://doi.org/10.18608/hla17.029). URL: <https://solaresearch.org/hla-17/hla17-chapter29>.

# Abreviaciones

**ANII** Agencia Nacional de Investigación e Innovación. 74, 75

**CC-BY** Creative Commons Attribution License. 29, 49

**DOI** Digital Object Identifier. 50

**FIng** Facultad de Ingeniería. 2, 27

**HTML** Hypertext Markup Language. 10, 47

**HTTP** HyperText Transfer Protocol. 15, 63

**InCo** Instituto de Computación. 2, 3, 27, 47

**JSON** JavaScript Object Notation. 10

**OCLC** Online Computer Library Center. 50

**ORCID** Open Researcher and Contributor ID. 50

**OWL** Web Ontology Language. 6, 9, 11, 16

**RDF** Resource Description Framework. 2, 6, 9, 10, 15, 37, 53, 85

**RDFS** Resource Description Framework Schema. 6, 9, 11, 16

**RIF** Rule Interchange Format. 7



**SNI** Sistema Nacional de Investigadores. 55, 75

**SPARQL** SPARQL Protocol and RDF Query Language. 2, 7, 11, 15, 94

**SQL** Structured Query Language. 11

**SWRL** Semantic Web Rule Language. 7

**UdelaR** Universidad de la República. 2, 27

**URI** Uniform Resource Identifiers. 6, 15, 36, 49

**W3C** World Wide Web Consortium. 2, 5–7, 11, 18

**XML** eXtensible Markup Language. 6, 8, 10

**XQuery** XML Query. 11

**XSD** XML Schema Definition. 8

**XSLT** Extensible Stylesheet Language Transformations. 37

# Nomenclatura

**API** (Application Programming Interface) es un conjunto de subrutinas, protocolos de comunicación y herramientas para desarrollar software. En términos generales define una forma clara de comunicación entre distintos componentes de la aplicación. [24](#), [32](#), [46](#)

**BibTeX** este término es utilizado para hacer mención a una herramienta y un formato de archivo que se utilizan para describir y procesar listas de referencias, principalmente en conjunto con documentos LaTeX. [47](#), [51](#)

**CNXML** es el que se almacenan los módulos, es decir, los contenidos propiamente dichos de la plataforma CNX. Esta basado en XML y se compone por etiquetas específicas como: *definition, meaning, term, exercise, problem, solution*, etc. [29](#)

**CollXML** es el formato en el que se almacenan las colecciones de CNX. Esta basado en XML y contiene la lista ordenada de módulos y sub-colecciones que componen un determinado libro de texto. [29](#)

**crawler** es un programa que inspecciona y procesa las páginas o documentos Web de forma recurrente y automatizada. Su objetivo es realizar análisis de distinto tipo fuera de línea y es comúnmente utilizado por motores de indexación y búsqueda. [24](#), [32](#), [33](#)

**EPUB** es un formato de archivo para almacenar libros electrónicos. Se pueden leer en distintos dispositivos que contengan el software adecuado, tales como teléfonos inteligentes, tabletas, computadoras o lectores electrónicos. [30](#)

**MDML** es el formato en el cual se almacenan los metadatos. Es utilizado en las

secciones de metadatos de las colecciones (CollXML) y módulos (CNXML).  
29

**OAI-PMH** es un protocolo creado por la *Open Archives Initiative* que está orientado a la extracción de metadatos. Funciona sobre el protocolo HTTP y describe como debe ser la interacción con el proveedor del servicio para el listado los nuevos registros y acceder a los metadatos. 35, 85

**ORDBMS** (Object-Relational Database Management System) es un manejador de base de datos híbrido entre el modelo orientado a objetos (OODBMS) y el modelo relacional (RDBMS) . 53

**precisión** (también conocido como el valor predictivo positivo) es la fracción de instancias relevantes del total de instancias recuperadas. En nuestro contexto se entiende como la cantidad de entidades correctamente identificadas del total de entidades enlazadas, es decir, las entidades enlazadas sin contar los falsos positivos. 4, 43

**RDBMS** (Relational Database Management System ORDBMS) es un tipo particular de manejador de base de datos que utiliza un modelo relacional para representar sus bases de datos, por lo tanto permite crear bases de datos relacionales. 53

**RSS** (Really Simple Syndication) es un formato XML para distribuir contenido en la web. Se utiliza para difundir información actualizada y novedades a usuarios que se han suscrito a la fuente de contenidos. 32

**sensibilidad** también conocida como recuperación o *recall* es la fracción de instancias relevantes que han sido recuperadas del total de instancias existentes. En este contexto se entiende como la cantidad de entidades correctamente identificadas (o enlazadas) del total de instancias identificables en la fuente destino. 4, 43

**UNICODE** es un estándar de codificación de caracteres diseñado para facilitar el tratamiento informático, transmisión y visualización de textos de múltiples lenguajes. 6

**XSLT** (Extensible Stylesheet Language Transformations) es un lenguaje cuyo

principal propósito es transformar documentos XML en diferentes formatos, tales como XML, RDF, HTML y texto plano, entre otros. 85

# Apéndice A

## Caso CNX

Ejemplo A.1: Transformación **XSLT** para la generación de **RDF** a partir del formato `cnx_dc` del protocolo **OAI-PMH**.

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <!-- NOTA: - los patrones "http://baseURI/" y http://sourceURI/ se
3      reemplazan en tiempo de ejecucion
4      - se modifica transformacion para soportar metadatos del formato
5      cnx_dc de cnx.org -->
6 <xsl:stylesheet version="1.0 "
7   xmlns:xsl="http://www.w3.org/1999/XSL/Transform "
8   xmlns:dc="http://purl.org/dc/elements/1.1/"
9   xmlns:dcterms="http://purl.org/dc/terms/"
10  xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
11  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
12  xmlns:oai="http://www.openarchives.org/OAI/2.0/"
13  xmlns:oai_voc="http://www.mediaspaces.info/vocab/oai-pmh.rdf#"
14  xmlns:oai_cnx="http://www.openarchives.org/OAI/2.0/oai_cnx/"
15  xmlns:cnxdc="http://cnx.org/technology/schemas/cnx_dc/"
16  xmlns:ore="http://www.openarchives.org/ore/terms/"
17  xmlns:dbc="http://dbpedia.org/resource/Category:">
18 <xsl:output method="xml" indent="yes" />
19 <xsl:template match="/">
20   <rdf:RDF
21     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
22     xmlns:dc="http://purl.org/dc/elements/1.1/"
23     xmlns:dcterms="http://purl.org/dc/terms/"
24     xmlns:dbc="http://dbpedia.org/resource/Category:"
25     xml:base="http://baseURI/">
26     <xsl:apply-templates select="//oai:record" />
27   </rdf:RDF>
28 </xsl:template>
29 <xsl:template match="//oai:record">
30   <oai_voc:Item>
31     <xsl:attribute name="rdf:about">
32       <xsl:value-of
```

```

31         select="concat('/resource/document/', oai:header/oai:identifier
32     )" />
33     </xsl:attribute>
34     <xsl:for-each select="oai:header/oai:setSpec">
35         <oai_voc:set>
36             <xsl:attribute name="rdf:resource">
37                 <xsl:value-of
38                     select="concat('/resource/set/', .)" />
39                 </xsl:attribute>
40             </oai_voc:set>
41         </xsl:for-each>
42
43         <oai_voc:origin>
44             <xsl:attribute name="rdf:resource">
45                 <xsl:value-of
46                     select="concat('http://sourceURI/', '?', 'verb=GetRecord', '&
47                     amp;', 'metadataPrefix=cnx_dc', '&', 'identifier=', oai:header/
48                     oai:identifier)" />
49                 </xsl:attribute>
50             </oai_voc:origin>
51             <xsl:apply-templates select="oai:metadata" />
52         </oai_voc:Item>
53     </xsl:template>
54     <xsl:template match="oai:metadata">
55         <xsl:for-each select="cnxdc:dc/dc:title">
56             <dc:title>
57                 <xsl:value-of select="." />
58             </dc:title>
59         </xsl:for-each>
60         <xsl:for-each select="cnxdc:dc/dc:creator">
61             <dc:creator>
62                 <xsl:value-of select="." />
63             </dc:creator>
64         </xsl:for-each>
65         <!-- tockeinsamos las palabras clave que estan aplanadas en CSV y
66         asignamos
67         dc:subject -->
68         <xsl:for-each select="cnxdc:dc/dc:subject">
69             <xsl:call-template name="splitText">
70                 <xsl:with-param name="csvText">
71                     <xsl:value-of select="." />
72                 </xsl:with-param>
73                 <xsl:with-param name="element" select="'dc:subject'" />
74             </xsl:call-template>
75         </xsl:for-each>
76         <xsl:for-each select="cnxdc:dc/dc:description">
77             <dc:description>
78                 <xsl:value-of select="." />
79             </dc:description>
80         </xsl:for-each>
81         <xsl:for-each select="cnxdc:dc/dc:publisher">
82             <dc:publisher>
83                 <xsl:value-of select="." />
84             </dc:publisher>
85         </xsl:for-each>
86         <xsl:for-each select="cnxdc:dc/dc:contributor">
87             <dc:contributor>
88                 <xsl:value-of select="." />
89             </dc:contributor>
90         </xsl:for-each>

```

```

89 <xsl:for-each select="cnxdc:dc/dc:date">
90 <dc:date>
91 <xsl:value-of select="." />
92 </dc:date>
93 </xsl:for-each>
94 <xsl:for-each select="cnxdc:dc/dc:type">
95 <dc:type>
96 <xsl:value-of select="." />
97 </dc:type>
98 </xsl:for-each>
99 <xsl:for-each select="cnxdc:dc/dc:format">
100 <dc:format>
101 <xsl:value-of select="." />
102 </dc:format>
103 </xsl:for-each>
104 <xsl:for-each select="cnxdc:dc/dc:identifier">
105 <xsl:choose>
106 <xsl:when test="starts-with(text(),'http://')">
107 <dc:identifier>
108 <xsl:attribute name="rdf:resource">
109 <xsl:value-of select="text()" />
110 </xsl:attribute>
111 </dc:identifier>
112 </xsl:when>
113 <xsl:otherwise>
114 <dc:identifier>
115 <xsl:value-of select="text()" />
116 </dc:identifier>
117 </xsl:otherwise>
118 </xsl:choose>
119 </xsl:for-each>
120 <xsl:for-each select="cnxdc:dc/dc:source">
121 <dc:source>
122 <xsl:value-of select="." />
123 </dc:source>
124 </xsl:for-each>
125 <xsl:for-each select="cnxdc:dc/dc:language">
126 <dc:language>
127 <xsl:value-of select="." />
128 </dc:language>
129 </xsl:for-each>
130 <xsl:for-each select="cnxdc:dc/dc:relation">
131 <dc:relation>
132 <xsl:value-of select="." />
133 </dc:relation>
134 </xsl:for-each>
135 <xsl:for-each select="cnxdc:dc/dc:coverage">
136 <dc:coverage>
137 <xsl:value-of select="." />
138 </dc:coverage>
139 </xsl:for-each>
140 <xsl:for-each select="cnxdc:dc/dc:rights">
141 <dc:rights>
142 <xsl:value-of select="." />
143 </dc:rights>
144 </xsl:for-each>
145 <!-- Tratamos el main subject de forma especial usamos
dcterms:subject -->
146 <xsl:for-each select="cnxdc:dc/cnxdc:cnx-subject">
147 <xsl:call-template name="splitSubject">
148 <xsl:with-param name="csvText">
149 <xsl:value-of select="." />

```

```

150         </xsl:with-param>
151         <xsl:with-param name="element"
152             select="'dcterms:subject'" />
153     </xsl:call-template>
154 </xsl:for-each>
155 <!-- Tratamos a los mantenedores como dc:contributor -->
156 <xsl:for-each select="cnxdc:dc/cnxdc:maintainer">
157     <dc:contributor>
158         <xsl:value-of select="." />
159     </dc:contributor>
160 </xsl:for-each>
161 </xsl:template>
162 <!-- template para tokenizar elementos aplanados a CSV -->
163 <xsl:template match="text()" name="splitText">
164     <xsl:param name="csvText" />
165     <xsl:param name="element" />
166
167     <xsl:choose>
168         <xsl:when test="contains($csvText,',' )">
169             <xsl:call-template name="splitText">
170                 <xsl:with-param name="csvText"
171                     select="substring-before($csvText,',' )" />
172                 <xsl:with-param name="element" select="$element" />
173             </xsl:call-template>
174             <xsl:call-template name="splitText">
175                 <xsl:with-param name="csvText"
176                     select="substring-after($csvText,',' )" />
177                 <xsl:with-param name="element" select="$element" />
178             </xsl:call-template>
179         </xsl:when>
180         <xsl:otherwise>
181             <xsl:element name="{ $element }">
182                 <xsl:value-of select="$csvText" />
183             </xsl:element>
184         </xsl:otherwise>
185     </xsl:choose>
186 </xsl:template>
187 <!-- template para tokenizar elementos aplanados a CSV, ademas
188     enlaza con
189     categoria fija de dblp. Para ser usado con el mail-subject -->
190 <xsl:template match="text()" name="splitSubject">
191     <xsl:param name="csvText" />
192     <xsl:param name="element" />
193     <xsl:choose>
194         <xsl:when test="contains($csvText,',' )">
195             <xsl:call-template name="splitSubject">
196                 <xsl:with-param name="csvText"
197                     select="substring-before($csvText,',' )" />
198                 <xsl:with-param name="element" select="$element" />
199             </xsl:call-template>
200             <xsl:call-template name="splitSubject">
201                 <xsl:with-param name="csvText"
202                     select="substring-after($csvText,',' )" />
203                 <xsl:with-param name="element" select="$element" />
204             </xsl:call-template>
205         </xsl:when>
206         <xsl:otherwise>
207             <xsl:choose>
208                 <xsl:when test="$csvText='Science and Technology'">
209                     <xsl:element name="{ $element }">
210                         <xsl:attribute name="rdf:resource">dbc:Science</
211 xsl:attribute>

```



```

210         </xsl:element>
211         <xsl:element name="{ $element }">
212             <xsl:attribute name="rdf:resource">dbc:Technology </
xsl:attribute>
213         </xsl:element>
214     </xsl:when>
215     <xsl:when test="$csvText='Mathematics and Statistics' ">
216         <xsl:element name="{ $element }">
217             <xsl:attribute name="rdf:resource">dbc:Mathematics</
xsl:attribute>
218         </xsl:element>
219         <xsl:element name="{ $element }">
220             <xsl:attribute name="rdf:resource">dbc:Statistics</
xsl:attribute>
221         </xsl:element>
222     </xsl:when>
223     <xsl:when test="$csvText='Social Sciences' ">
224         <xsl:element name="{ $element }">
225             <xsl:attribute name="rdf:resource">dbc:Social_sciences</
xsl:attribute>
226         </xsl:element>
227     </xsl:when>
228     <xsl:when test="$csvText='Humanities' ">
229         <xsl:element name="{ $element }">
230             <xsl:attribute name="rdf:resource">dbc:Humanities</
xsl:attribute>
231         </xsl:element>
232     </xsl:when>
233     <xsl:when test="$csvText='Arts' ">
234         <xsl:element name="{ $element }">
235             <xsl:attribute name="rdf:resource">dbc:Arts</
xsl:attribute>
236         </xsl:element>
237     </xsl:when>
238     <xsl:when test="$csvText='Business' ">
239         <xsl:element name="{ $element }">
240             <xsl:attribute name="rdf:resource">dbc:Business</
xsl:attribute>
241         </xsl:element>
242     </xsl:when>
243     <xsl:otherwise>
244         <xsl:element name="{ $element }">
245             <xsl:value-of select="$csvText" />
246         </xsl:element>
247     </xsl:otherwise>
248 </xsl:choose>
249 </xsl:otherwise>
250 </xsl:choose>
251 </xsl:template>
252 </xsl:stylesheet>

```

# Apéndice B

## Caso InCo

Ejemplo B.1: Vocabulario en formato Turtle creado para describir a un docente. Reutiliza propiedades del vocabulario FOAF y DBpedia. A su vez define nuevas propiedades que describen la relación laboral con la institución.

```
1  @prefix : <http://data.fing.edu.uy/schema/docentes/> .
2  @prefix owl: <http://www.w3.org/2002/07/owl#> .
3  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
4  @prefix xml: <http://www.w3.org/XML/1998/namespace> .
5  @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
6  @prefix foaf: <http://xmlns.com/foaf/0.1/> .
7  @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
8  @prefix docentes: <http://data.fing.edu.uy/schema/docentes/> .
9  @base <http://data.fing.edu.uy/schema/docentes/> .
10
11 <http://data.fing.edu.uy/schema/docentes/> rdf:type owl:Ontology ;
12     owl:imports <http://
13         dbpedia.org/ontology/> ,
14         foaf: .
15
16 #####
17 # Datatypes
18 #####
19
20 ### http://data.fing.edu.uy/schema/docentes/grado
21 docentes:grado rdf:type rdfs:Datatype ;
22     owl:equivalentClass [ rdf:type rdfs:Datatype ;
23         owl:onDatatype xsd:int ;
24         owl:withRestrictions ( [ xsd:
25             minInclusive "1"^^xsd:int
26                                     ]
27             [ xsd:
28                 maxInclusive "5"^^xsd:int
29                                     ]
30         )
31     ] .
```

29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78

```
#####  
# Object Properties  
#####  
  
### http://xmlns.com/foaf/0.1/homepage  
foaf:homepage rdfs:domain docentes:Docente .  
  
### http://xmlns.com/foaf/0.1/phone  
foaf:phone rdfs:domain docentes:Docente .  
  
#####  
# Data properties  
#####  
  
### http://data.fing.edu.uy/schema/docentes/RDT  
docentes:RDT rdf:type owl:DatatypeProperty ;  
rdfs:domain docentes:Docente ;  
rdfs:range xsd:boolean ;  
<http://purl.org/dc/elements/1.1/description> "Indica si  
el docente esta en regimen de deficación total"@es ,  
"Régimen de  
Dedicación Total (RDT): Este régimen conlleva la dedicación  
exclusiva , y otorga una compensación salarial del 60% sobre el  
sueldo base. Los docentes en RDT pueden realizar tareas puntuales  
relacionadas con su especialidad , para lo cual deben solicitar  
autorización previa al servicio universitario en el que se desempe  
ñan."@es ;  
rdfs:isDefinedBy <http://data.fing.edu.uy/schema/docentes  
> ;  
rdfs:label "RDT" ;  
rdfs:seeAlso "http://dedicaciontotal.udelar.edu.uy/"@es .  
  
### http://data.fing.edu.uy/schema/docentes/grado  
docentes:grado rdf:type owl:DatatypeProperty ;  
rdfs:domain docentes:Docente ;  
rdfs:range docentes:grado .  
  
### http://data.fing.edu.uy/schema/docentes/horas  
docentes:horas rdf:type owl:DatatypeProperty ;  
rdfs:domain docentes:Docente ;  
rdfs:range xsd:int ;  
<http://purl.org/dc/elements/1.1/description> "Cantidad  
de horas que trabaja en la institución"@es ;  
rdfs:isDefinedBy <http://data.fing.edu.uy/schema/  
docentes/> ;  
rdfs:label "horas" .  
  
### http://data.fing.edu.uy/schema/docentes/oficina  
docentes:oficina rdf:type owl:DatatypeProperty ;  
rdfs:domain docentes:Docente ;  
rdfs:range xsd:int ;  
<http://purl.org/dc/elements/1.1/description> "Número  
de oficina dentro del Instituto."@es ;  
rdfs:isDefinedBy <http://data.fing.edu.uy/schema/  
docentes/> ;
```

```

79         rdfs:label "oficina" .
80
81
82     ### http://dbpedia.org/ontology/orcidId
83     <http://dbpedia.org/ontology/orcidId> rdf:type owl:DatatypeProperty ;
84         rdfs:domain docentes:Docente ;
85         rdfs:label "ORCID Id"@en .
86
87
88     ### http://xmlns.com/foaf/0.1/name
89     foaf:name rdfs:domain docentes:Docente .
90
91
92     #####
93     # Classes
94     #####
95
96     ### http://data.fing.edu.uy/schema/docentes/Docente
97     docentes:Docente rdf:type owl:Class ;
98         rdfs:subClassOf foaf:Person ;
99         <http://purl.org/dc/elements/1.1/description> "Se
100     denominan docentes las personas que ocupan cargos docentes en
101     efectividad o en forma interina o que desempeñan funciones
102     docentes en calidad de docentes contratados, honorarios o libres.
103     Art. 5 del Estatuto Personal Docente de 15/04/1968. Los cargos
104     docentes se agrupan en 5 grados identificados por orden jerárquico
105     creciente." ;
106     rdfs:isDefinedBy <http://data.fing.edu.uy/schema/
107     docentes/> ;
108         rdfs:label "Docente"@es ;
109         rdfs:seeAlso "http://gestion.udelar.edu.uy/
110     planeamiento/funcionarios/funcionarios-docentes/"@es .
111
112
113     ### http://xmlns.com/foaf/0.1/Agent
114     foaf:Agent rdf:type owl:Class .
115
116
117     #####
118     # Annotations
119     #####
120
121     docentes:grado rdfs:seeAlso "http://dgp.udelar.edu.uy/renderPage/index
122     /pageId/699"@es ,
123         "http://gestion.udelar.edu.uy/planeamiento
124     /wp-content/uploads/sites/27/2013/04/estatuto_personal_docente.pdf
125     "@es ;
126         <http://purl.org/dc/elements/1.1/description> "Escalafó
127     n G – Docente: Agrupa a los cargos docentes que cumplen funciones
128     establecidas en la Ordenanza de Organización Docente. Los grados
129     equivalen a: 1–Ayudante, 2–Asistente, 3–Profesor Adjunto, 4–
130     Profesor Agregado, 5–Profesor Titular."@es ;
131     rdfs:isDefinedBy <http://data.fing.edu.uy/schema/
132     docentes/> ;
133         rdfs:label "grado" .
134
135
136     ### Generated by the OWL API (version 4.2.8.20170104-2310) https://
137     github.com/owlcs/owlapi

```

Ejemplo B.2: Consulta SPARQL que obtiene la cantidad de publicaciones, coautorías y el grado de cada docente del InCo. Esta consulta se utilizó como base para varios de los gráficos presentados, integra datos de las tres fuentes mencionadas utilizando los enlaces detectados en las etapas anteriores.

```

1 #####
2 # Cantidad de publicaciones de biblio y de dblp, coautores,
3 # y coautores InCo por docente, grado
4 #####
5 PREFIX docente: <http://data.fing.edu.uy/schema/docentes/>
6 PREFIX dc: <http://purl.org/dc/elements/1.1/>
7 PREFIX owl: <http://www.w3.org/2002/07/owl#>
8 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
9
10 SELECT
11   ?autor_inco
12   ?autor_inco_nombre
13   ?grado
14   (COUNT( DISTINCT ?doc_biblio) AS ?cant_pub_biblio)
15   (COUNT( DISTINCT ?doc) AS ?cant_pub)
16   (COUNT( DISTINCT ?coautor_inco) AS ?cant_coautores_inco)
17   (COUNT( DISTINCT ?coautor_dblp) AS ?cant_coautores)
18 #grafos con datos
19 FROM NAMED <http://data.fing.edu.uy/docentes#>
20 FROM NAMED <http://data.fing.edu.uy/biblio#>
21 FROM NAMED <http://dblp.org>
22 #enlaces generados
23 FROM NAMED <http://data.fing.edu.uy/links/inco_dblp#>
24 FROM NAMED <http://data.fing.edu.uy/links/inco_biblio#>
25 FROM NAMED <http://data.fing.edu.uy/links/biblio_dblp#>
26 WHERE
27 {
28   #partimos de los docentes enlazados
29   GRAPH <http://data.fing.edu.uy/links/inco_dblp#> {
30     ?autor_inco owl:sameAs ?autor_dblp
31     . filter (?autor_inco!=<http://data.fing.edu.uy/resource/docente/
32     eduardof>)
33   }
34   #datos extra del docente
35   GRAPH <http://data.fing.edu.uy/docentes#> {
36     ?autor_inco foaf:name ?autor_inco_nombre .
37     OPTIONAL {
38       ?autor_inco docente:grado ?grado
39     }
40   }
41   #datos extra la publicación
42   GRAPH <http://dblp.org> {
43     ?doc a foaf:Document;
44     dc:title ?titulo;
45     foaf:maker ?autor_dblp .
46     OPTIONAL {
47       #opcionalmente traemos coautores (podía no tener), excluimos al
48       autor de la lista de coautores
49       ?doc foaf:maker ?coautor_dblp . FILTER(?autor_dblp != ?
50       coautor_dblp) .
51     }
52     OPTIONAL {
53       #opcionalmente tiene un tipo de documento (lo hacemos
54       #de esta forma pues la prop rdf:label esta en alemán)

```

```

52     ?doc a ?class .
53     FILTER(?class!=foaf:Document) .
54     BIND(STRAFTER( xsd:string(?class), "#") as ?class_name)
55   }
56 } .
57 OPTIONAL {
58   #opcionalmente vemos cuales de esos coautores son posiblemente del
59   InCo
60   GRAPH <http://dblp.org> {
61     ?doc a foaf:Document;
62     foaf:maker ?autor_dblp ;
63     foaf:maker ?coautor_dblp . FILTER(?autor_dblp != ?
64     coautor_dblp)
65   }
66   GRAPH <http://data.fing.edu.uy/links/inco_dblp#> {
67     ?coautor_inco owl:sameAs ?coautor_dblp .
68   }
69 }
70 OPTIONAL {
71   #opcionalmente vemos si el docente tiene publicaciones en biblio
72   FIng
73   GRAPH <http://data.fing.edu.uy/links/inco_biblio#> {
74     ?autor_inco owl:sameAs ?autor_biblio
75   }
76   GRAPH <http://data.fing.edu.uy/biblio#> {
77     ?doc_biblio dc:creator ?autor_biblio ;
78     dc:title ?doc_biblio_titulo .
79   }
80 }
81 }
82 GROUP BY ?autor_inco ?grado ?autor_inco_nombre

```

En el Ejemplo B.2 se puede observar la consulta SPARQL de agrupación que combina datos de la fuente DBLP, *Docentes InCo* y *Biblio FIng*. En las líneas 5 a 8 se definen los prefijos de vocabularios que se van a utilizar dentro de la consulta, esto permite abreviar las consultas y hacerlas más legibles. De la línea 11 a 17 se define la proyección, es decir, las distintas variables que se van a retornar en el resultado. En este caso nos interesa el identificador, grado y nombre del docente. Además se agregaron otras variables que nos interesa contar en la agrupación. De las líneas 18 a 25 se especifica de qué grafos se van a sacar los datos. En este caso tienen la cláusula NAMED que indica que se hará referencia explícita por su nombre, lo que nos da mayor control sobre los patrones y filtros que se definan. Como se puede observar tenemos los grafos con los datos de las fuentes publicadas, así como los grafos con los enlaces detectados. Cabe destacar que en este caso se cuenta con todos los datos cargados localmente, inclusive DBLP, lo que nos permite hacer referencia de esta forma. De no ser así, el estándar SPARQL admite definir consultas federadas utilizando la cláusula SERVICE<sup>114</sup>. De ser necesario se podrían reemplazar las

<sup>114</sup><https://www.w3.org/TR/sparql11-federated-query/>

cláusulas `GRAPH <http://dblp.org> {...}` por cláusulas `SERVICE` del estilo de `SERVICE <https://dblp.13s.de/d2r/sparql> {...}` para que dicha porción de la consulta se envíe a otro terminal SPARQL para su resolución. Luego se da paso a la sección `WHERE` donde se definen los distintos patrones, es decir, la estructura de grafo que deben cumplir las tripletas para ser tomadas en cuenta en el resultado. El vínculo entre los distintos patrones se realiza a través de las variables que se enlazan. Por ejemplo de la línea 29 a la 31 se define el punto de partida que asocia un docente del InCo y su correspondiente autor DBLP del grafo de enlaces detectados a la variable `?autor_inco` y `?autor_dblp` respectivamente. Estas variables se irán utilizando en otros patrones para definir otras variables y filtros. Los bloques `GRAPH` indican que los patrones se deben verificar contra el grafo especificado. Por ejemplo, de la línea 33 a la 38 se van a buscar más datos del docente al grafo correspondiente. En este caso se trae el nombre y opcionalmente el grado. El uso de la cláusula `OPTIONAL` nos permite especificar un patrón que debe cumplirse opcionalmente. Si el patrón no fuera opcional en el ejemplo, se excluiría del resultado a todos los docentes que no tengan grado. En una web de fuentes independientes esta cláusula es de mucha utilidad ya que nos permite lidiar con datos parciales, lo cual es muy común en la web de datos. La cláusula `FILTER` permite especificar expresiones booleanas que deben ser verdaderas para ser incluidas en la solución, mientras que la cláusula `BIND` permite asociar un valor a una variable, por ejemplo proveniente de una manipulación de cadena de caracteres como el caso en la línea 53. Finalmente en la última línea se define la cláusula de agrupación.

Ejemplo B.3: Consulta SPARQL que obtiene la lista de 20 conferencias con mayor cantidad publicaciones de docentes del InCo en el periodo 2006-2016.

```

1  #####
2  #top 10 conferencias
3  #####
4  PREFIX docente: <http://data.fing.edu.uy/schema/docentes/>
5  PREFIX dct: <http://purl.org/dc/terms/>
6  PREFIX dc: <http://purl.org/dc/elements/1.1/>
7  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
8  PREFIX owl: <http://www.w3.org/2002/07/owl#>
9  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
10 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
11 PREFIX swrc: <http://swrc.ontoware.org/ontology#>
12 PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
13
14 SELECT
15     ?conf_type_name
16     ?conf
17     (SAMPLE(?conf_name ) as ?conf_name) #puede tener varios nombres
18     (COUNT(DISTINCT ?doc) as ?cant_docs)
19     (COUNT(DISTINCT ?autor_inco) as ?cant_docentes)
20 #grafos con datos
21 FROM NAMED <http://dblp.org>

```

```

22 #grafos con enlaces
23 FROM NAMED <http://data.fing.edu.uy/links/inco_dblp#>
24 WHERE {
25   GRAPH <http://data.fing.edu.uy/links/inco_dblp#> {
26     ?autor_inco owl:sameAs ?autor_dblp .
27   }
28   GRAPH <http://dblp.org> {
29     ?doc a foaf:Document;
30     foaf:maker ?autor_dblp;
31     swrc:series ?conf; #restringe a conferencias
32     dct:issued ?year .
33     FILTER(year(?year) >= 2006 && year(?year) <= 2016 )
34     ?conf a ?conf_type;
35     rdfs:label ?conf_name .
36     BIND(STRAFTER( xsd:string(?conf_type), "#") as ?conf_type_name)
37   }
38 }
39 GROUP BY ?conf ?conf_type_name
40 ORDER BY desc(?cant_docs)
41 LIMIT 20

```

En esta consulta se aplican los mismos conceptos descritos en el Ejemplo B.2. Partimos del grafo de enlaces de docentes del InCo con autores de DBLP en la línea 26. Una vez que se tienen los autores de DBLP buscamos los documentos en los que figuran como autores mediante la propiedad *foaf:maker* en la línea 30. Con la fecha de publicación definida con la propiedad *dct:issued* se aplica el filtro de rango de fechas de la línea 33. Finalmente se agrupa por conferencia, se ordena de forma descendente por cantidad de documentos distintos y nos quedamos con los primeros 20. Es importante destacar el patrón de la línea 31 donde se restringe a los documentos que tengan la propiedad *swrc:serie*. Dicha propiedad es parte de los recursos de tipo *swrd:InProceedings*, y es la que asocia al documento con la conferencia donde se publicó. De forma análoga pero utilizando la propiedad *swrc:journal* restringimos a documentos de tipo *swrc:Article* publicados en revistas, y de esta forma obtenemos los datos que complementan la Tabla 3.13. Al igual que en otros lenguajes de consulta es fundamental tener conocimiento sobre la estructura de los datos a la hora de crear las consultas. Gracias al uso de vocabularios abiertos y conocidos este conocimiento queda explícito y accesible junto con los datos.