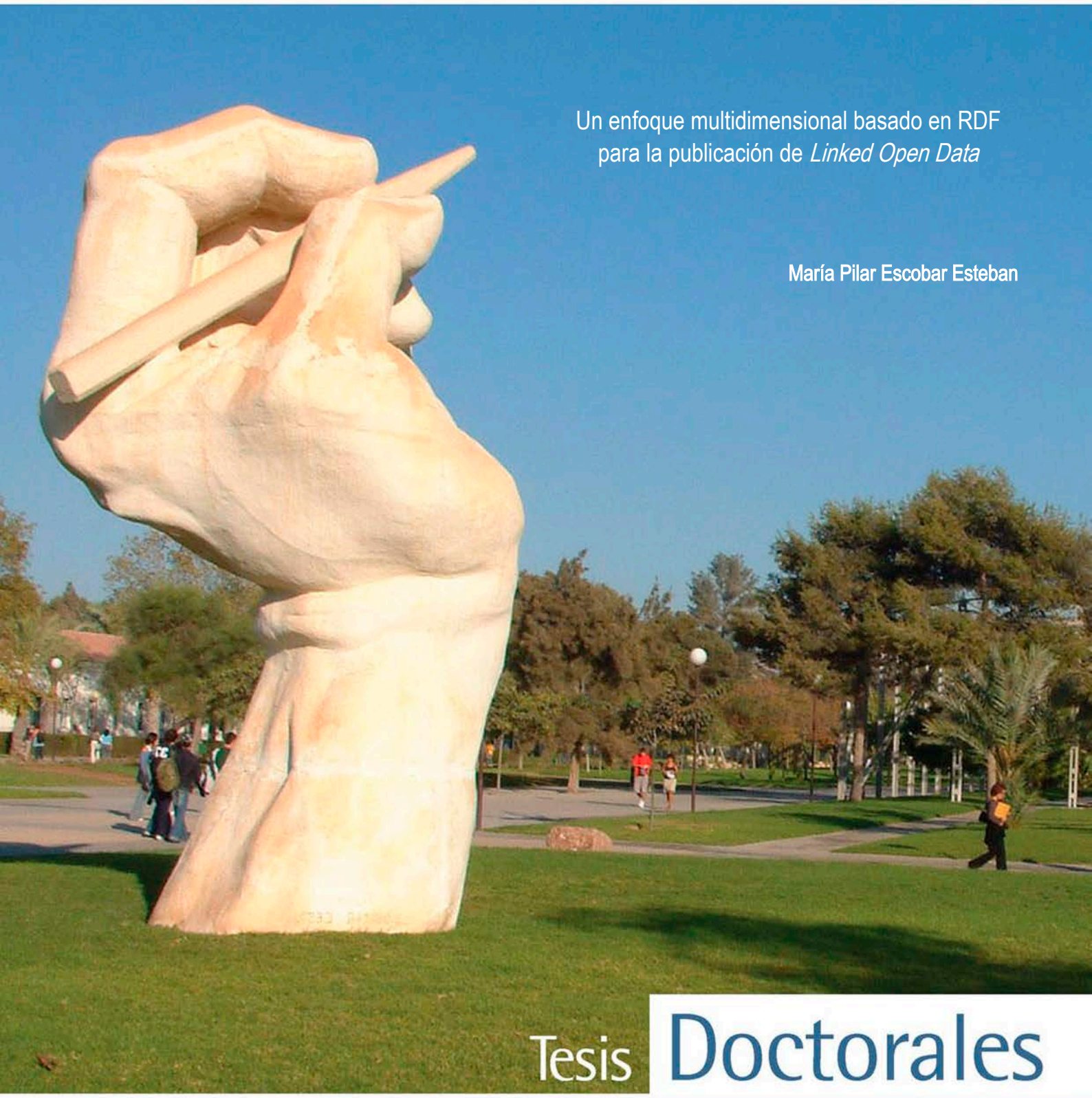
Universitat d'Alacant
Universidad de Alicante

Un enfoque multidimensional basado en RDF para la publicación de *Linked Open Data*

María Pilar Escobar Esteban



Tesis Doctorales
UNIVERSIDAD de ALICANTE

Unitat de Digitalització UA
Unidad de Digitalización UA

UNIVERSIDAD DE ALICANTE
DEPARTAMENTO DE LENGUAJES Y SISTEMAS INFORMÁTICOS

Un enfoque multidimensional basado en RDF para la publicación de *Linked Open Data*

María Pilar Escobar Esteban
TESIS DOCTORAL

Directores:
Dr. Jesús Peral Cortés
Dr. Manuel Marco Such

Julio de 2020

TESIS DOCTORAL

Un enfoque multidimensional basado en RDF para la publicación de *Linked Open Data*

Este documento contiene una síntesis del trabajo realizado en esta tesis por María Pilar Escobar Esteban, bajo la dirección de los doctores Jesús Peral Cortés y Manuel Marco Such, para optar al grado de Doctora en Informática.

Julio de 2020

Dedicatoria

A mi familia, por su apoyo incondicional, que siempre me animó a seguir adelante.

Agradecimientos

En primer lugar, quiero agradecer a mis directores de tesis doctoral Manuel Marco Such y Jesús Peral Cortés, por todo el apoyo brindado y los consejos que me han proporcionado en toda mi trayectoria investigadora. Todo ello me ha permitido realizar distintas publicaciones en congresos y revistas de impacto que han culminado con la presentación de este trabajo. A Gustavo Candela y María Dolores Sáez, por su apoyo y colaboración en todo momento. Su colaboración ha sido fundamental para la consecución de este trabajo. También quiero agradecer a mis compañeros de la Biblioteca Virtual Miguel de Cervantes; a Rafael Carrasco por su inestimable orientación en la investigación, a Jesús Pradells, su director, por su colaboración y apoyo en todos estos años de actividad investigadora; a Pedro Pernías, porque con él empezamos toda esta aventura de la Biblioteca Virtual. A Manuel Bravo por su apoyo y facilidades en el desempeño de mi trabajo; a María García y a Fina Carrión por lo fácil que resulta trabajar con ellas. A Leonel Iriarte y Alexander Sánchez por todos sus consejos. Del mismo modo, con profunda sinceridad, agradezco a los compañeros del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante su ayuda tanto en tareas docentes como investigadoras. A Juan Carlos Trujillo y a su grupo de investigación LUCENTIA por ofrecerme la oportunidad de participar en el proyecto ECLIPSE y por su aportación.

Por último, y no por ello menos importante, agradezco a mi familia su apoyo y dedicación en estas etapas educativas e investigadoras que he desarrollado en la Universidad de Alicante.

Este trabajo ha sido parcialmente financiado por el Ministerio de

Ciencia, Innovación y Universidades de España a través del Proyecto ECLIPSE-UA (Enhancing Data Quality and Security for Improving Business Processes and Strategic Decisions in Cyber Physical Systems) con referencia RTI2018-094283-B-C32.

Índice general

Índice de figuras

Índice de tablas

Lista de acrónimos

API *Application Programming Interface* 4, 24, 31, 41, 55

BPM *Business Process Management* 46

BPMN *Business Process Model and Notation* 46, 154

BVMC Biblioteca Virtual Miguel de Cervantes 214

CSV *Comma-Separated Values* 12, 23, 30, 50, 52, 64, 65, 85, 86, 89, 90, 92, 95, 100

DCMI *Dublin Core Metadata Initiative* 37

ETL *Extract, Transform and Load* 49, 64, 65, 88, 89

FOAF *Friend of a Friend* 37

HTML *HyperText Markup Language* 104, 125, 128

HTTP *Hypertext Transfer Protocol* 26, 71, 99, 123

IA Inteligencia Artificial 153

IoT *Internet of Things* 2, 18

JCR *Journal Citation Reports* xxii, 155, 158

JSON *JavaScript Object Notation* 23, 26, 50, 64, 100

KG *Knowledge Graph* xxi, 3, 6, 28, 33, 56, 58, 83, 103, 150

LOD *Linked Open Data* ix, x, xxi, 1, 3, 4, 5, 6, 7, 9, 15, 16, 19, 28, 29, 30, 33, 37, 38, 39, 42, 43, 45, 47, 50, 52, 55, 58, 59, 61, 62, 63, 64, 68, 71, 75, 80, 83, 84, 89, 92, 93, 97, 102, 109, 120, 123, 149, 150, 151, 152, 153, 154, 156, 158, 214

MDA *Model-Driven Architecture* 154

MELODA *Methodology for releasing Open Data* xiii, 31, 35

OCR *Optical Character Recognition* 12

OWL *Web Ontology Language* 27, 67, 110, 179

PDF *Portable Document Format* 12, 24, 52, 86, 88, 92

RDF *Resource Description Framework* ix, x, 1, 17, 19, 23, 26, 27, 31, 36, 37, 38, 39, 40, 41, 42, 46, 50, 51, 53, 54, 56, 57, 59, 61, 64, 65, 68, 70, 72, 89, 91, 93, 94, 95, 97, 99, 100, 104, 106, 119, 121, 122, 123, 124, 125, 128, 131, 151, 179, 197, 200, 201, 204

RDFS *Resource Description Framework Schema* 27, 57

SCO *Smart City Ontology* 21, 25

SCOVO *The Statistical Core Vocabulary* 36

SDMX *Statistical Data and Metadata Exchange* 37

SKOS *Simple Knowledge Organization System* 36, 38, 122

SPARQL *Simple Protocol y RDF Query Language* x, xxii, 4, 26, 27, 37, 39, 40, 41, 45, 48, 56, 57, 60, 68, 75, 76, 75, 76, 80, 99, 102, 104, 111, 121, 124, 126, 128, 131, 132, 133, 141, 147, 150, 151

TIC *Tecnología de la Información y Comunicación* 39, 59, 131, 154

UE *Unión Europea* 84

URI *Uniform Resource Identifier* 1, 26, 27, 31, 37, 42, 51, 53, 56, 58, 65, 67, 71, 90, 93, 94, 97, 99, 102, 104, 118, 119, 128

VoID *Vocabulary of Interlinked Datasets* 57, 58, 117, 126

W3C *World Wide Web Consortium* 24, 26, 27, 30, 31, 35, 36, 40, 52, 64, 72, 91, 149, 151

XLS *Excel Spreadsheet* 11, 30, 64

XML *Extensible Markup Language* 23, 41, 50, 64

Resumen del trabajo realizado

Cada vez hay disponibles más datos de manera pública en Internet y surgen nuevas bases de conocimiento conocidas como *Knowledge Graph* (KG), basadas en conceptos de *Linked Open Data* (datos abiertos enlazados), como DBPedia, Wikidata, YAGO o *Google Knowledge Graph*, que cubren un amplio abanico de campos del conocimiento. Además, se incorporan los datos que provienen de diversas fuentes como dispositivos inteligentes o las redes sociales. Sin embargo, que estos datos estén públicos y accesibles no garantiza que sean útiles para los usuarios, no siempre se garantiza que sean confiables ni que puedan ser reutilizados de manera eficiente. Actualmente, siguen existiendo barreras que dificultan la reutilización de los datos, porque los formatos son poco adecuados para el procesamiento automático y publicación de la información, por falta de metadatos descriptivos y de semántica, duplicidades, ambigüedad o incluso errores en los propios datos. A todos estos problemas hay que añadir la complejidad del proceso de explotación de la información de un repositorio LOD. El trabajo y conocimientos técnicos que requiere el acceso, recolección, normalización y preparación de los datos para que puedan ser reutilizados supone una carga extra para los usuarios y organizaciones que quieran utilizarlos.

Para garantizar una eficiente explotación de los mismos, resulta fundamental dotarlos de más valor estableciendo conexiones con otros repositorios que permitan enriquecerlos; garantizar su valor, evaluando y mejorando la calidad de lo que se publica; y asimismo ofrecer los mecanismos necesarios que faciliten su explotación.

En este trabajo de tesis se ha propuesto un modelo para la pu-

blicación de LOD que, a partir de un conjunto de datos obtenidos de diversas fuentes, facilita la publicación, enriquecimiento y validación de LOD, generando información útil y de calidad orientada a usuarios expertos y no expertos. Las principales aportaciones son:

- La definición de un modelo para la publicación de datos abiertos enlazados con un enfoque multidimensional basado en RDF Data Cube, que garantiza la confiabilidad en los datos, mejorando la explotación y promoviendo su reutilización efectiva.
- El enriquecimiento de datos mediante repositorios externos como Wikidata y GeoNames.
- La incorporación de una metodología para la evaluación de la calidad del conjunto de datos basado en el vocabulario RDF Data Cube.
- La publicación de cuadros de mando que permiten utilizar y comprender de manera efectiva los datos generados de acuerdo al vocabulario RDF Data Cube. Además, se suministra un punto de acceso SPARQL público para que tanto las aplicaciones como usuarios expertos puedan conectar, consultar y recuperar la información contenida en el repositorio.

Como resultado de este trabajo de investigación se han publicado tres artículos en revistas indexadas en *Journal Citation Reports* (JCR). Además, se han realizado diversas actividades y comunicaciones para la difusión y fomento de los datos abiertos enlazados, así como otras publicaciones y presentaciones en las que se ha difundido la aplicación de estas metodologías aplicadas a otros dominios.

1. Introducción

> No importa lo lento que vayas mientras no te detengas.
>
> *Confucio*

En este capítulo se presentan los aspectos más relevantes relacionados con el desarrollo de LOD y cómo ha evolucionado a lo largo de los últimos años. Además, se describe la motivación, objetivos y estructura de esta tesis.

1.1 Propósito de este trabajo

Los datos son abiertos si son accesibles y están disponibles con las características tanto técnicas como legales necesarias para que se puedan usar, reutilizar y redistribuir libremente para cualquier propósito, sin ningún tipo de restricción [The World Bank, 2013]. La publicación de datos abiertos ha despertado un gran interés en diferentes dominios (gubernamental, académico, patrimonio cultural, etc.) pero los datos son accesibles y pueden utilizarse siempre y cuando estén publicados en un formato que permita su lectura, procesamiento automático y reutilización.

Tim Berners-Lee propuso una clasificación para medir la calidad de los datos que se publican basándose en su nivel de reutilización. Esta

clasificación es conocida como el esquema de desarrollo en 5 estrellas[1] y establece 5 niveles en función de la facilidad con la que puedan ser reutilizados. El nivel más bajo, una estrella, representa los datos que se publican bajo licencia abierta pero en cualquier formato independientemente de que sea difícil de manipular y reutilizar; es la forma más primitiva de publicar datos. En el siguiente nivel, dos estrellas, es cuando se publican como datos estructurados pero el formato es propietario, se pueden procesar y sigue siendo sencillo publicarlos. En el nivel tres estrellas, los datos son estructurados y en un formato abierto. El nivel cuatro estrellas representa el uso de *Uniform Resource Identifier* (URI) para identificar los recursos; implica un salto cuantitativo en cuanto a la reutilización y publicación de datos abiertos, pero requiere más esfuerzo para publicar los datos. Y finalmente la clasificación más alta, cinco estrellas, dada cuando los datos se enlazan con otros repositorios ofreciendo los datos enriquecidos al usuario, son los *Linked Open Data*, término acuñado por Tim Berner-Lee y presentado en TED2009[2]. Basados en RDF como modelo de datos para describir recursos, codifican la información en tripletas (sujeto, predicado y objeto) e identifican los recursos mediante URIs. Es la base para publicar, conectar e intercambiar datos.

Muchos son los productores de datos que se encuentran en las tres primeras posiciones de la clasificación cinco estrellas propuesta por Tim Berners-Lee, es decir, publican sus datos bajo licencia abierta y dependiendo del nivel de clasificación ofrecen datos estructurados y en formato libre. Según los datos recogidos por `datos.gob.es` aproximadamente el 80 % de las iniciativas identificadas se situaría en el tercer nivel de la clasificación (tres estrellas) y el 7 % en los dos primeros niveles.

[1] `https://5stardata.info/en/`

[2] `https://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide`

Cada vez se generan más datos, ciudades tan destacadas a nivel mundial como Barcelona, Madrid, Zaragoza, París y Nueva York se han convertido en grandes productoras de datos y ponen a disposición de los usuarios datos relativos a limpieza, consumos, agua, presupuestos, etc. Además, cada vez más personas están interconectadas, por ejemplo, en junio de 2018 más de 4 mil millones de usuarios en todo el mundo (aproximadamente el 55 % de la población mundial) estaban conectados a Internet. Se suman diversas fuentes de datos provenientes de distintos dispositivos inteligentes, sensores *Internet of Things* (IoT), redes sociales, etc. que arrojan a Internet una cantidad ingente de datos cada minuto. Organismos tan relevantes dentro del patrimonio cultural como la Biblioteca del Congreso de los Estados Unidos[3], la Biblioteca Británica[4], la Biblioteca Nacional Francesa[5], la Biblioteca Nacional de España[6] o la Biblioteca Virtual Miguel de Cervantes[7] publican sus repositorios de datos abiertos enlazados.

Además, están surgiendo varios grafos de conocimiento conocidos como KG, que son bases de conocimiento estructuradas basadas en conceptos de LOD; algunos ejemplos son Dbpedia[8], Wikidata[9], YAGO [Rebele et al., 2016] o incluso Google Knowledge Graph[10]. Estos KG son una fuente rica de conocimiento con la que enriquecer los datos, dado que cubren un amplio abanico general de conocimiento. Proporcionan datos estructurados, conectados, con acceso a descripciones y propiedades en varios idiomas entre otras.

Este escenario genera un gran volumen y variedad de datos que

[3] http://id.loc.gov/
[4] https://bnb.data.bl.uk/
[5] https://data.bnf.fr/
[6] http://datos.bne.es
[7] http://data.cervantesvirtual.com
[8] http://es.dbpedia.org/
[9] https://www.wikidata.org/
[10] https://developers.google.com/knowledge-graph

podrían convertirse en información útil para las instituciones, comunidad académica, ciudadanos, etc. Pero según un informe publicado por la Comisión Europea sobre la reutilización de *Open Data* [Portal, 2017], todavía quedan barreras internas y externas por superar que dificultan la expansión y reutilización de los datos abiertos. Pese a la existencia de esta gran cantidad de datos disponibles en Internet de diferentes dominios y del gran número de fuentes de información, a los expertos y usuarios de esos dominios les resulta complejo acceder y explotar esos datos. En general, los datos no son adecuados para los procesos ni computadoras tradicionales, a veces ni siquiera son procesables de manera automática, no siempre siguen los estándares y, además, se suman los problemas que se generan por la falta de semántica.

Por otro lado, está la complejidad existente en la reutilización de repositorios LOD por parte de usuarios no expertos. Aunque los datos abiertos enlazados se publiquen bajo una licencia abierta, consultarlos implica entender los conceptos semánticos contenidos en los repositorios así como tener conocimientos técnicos para acceder a la información. Muchos de estos repositorios suministran un punto de acceso SPARQL público o un *Application Programming Interface* (API) que requiere de conocimientos técnicos para consultar la información. Además, muchos de ellos no disponen de interfaces amigables que permitan la consulta de la información de manera sencilla. Actualmente sigue existiendo una brecha digital en lo que respecta a las herramientas y experiencia de acceso a los datos, sin olvidar que existe una parte de la sociedad que no tiene acceso a las tecnologías de información y comunicación careciendo de oportunidades de acceso a los datos.

Por todo lo mencionado anteriormente nos encontramos frente a dos grandes retos: (1) facilitar la publicación de datos abiertos a las instituciones garantizando que los datos sean accesibles, reutilizables y que cumplan con las exigencias de calidad requeridas; (2) suministrar nuevas interfaces de acceso a los repositorios que permitan que usuarios

no expertos puedan consultar la información disponible.

Resulta fundamental definir un método que facilite el proceso de publicación y acceso a los datos que se generan en las distintas instituciones tanto públicas como privadas, abriendo un nuevo escenario de colaboración entre los distintos actores, tanto productores como consumidores de datos. Y es que hoy en día, los datos son un factor clave tanto para la innovación social como para el crecimiento económico, sin olvidar que tienen un gran valor comercial [European Data Portal, 2015].

La integración de diferentes fuentes de datos, la interconexión entre estas bases de conocimientos y los datos generados por instituciones y ciudadanos, pueden suministrar una fuente muy rica de enriquecimiento semántico. Además, generando interfaces más sencillas se facilitaría el análisis de la información para los usuarios no expertos.

La finalidad de esta tesis es demostrar que se puede establecer un modelo que facilite la publicación de LOD, incluyendo en sus pasos el enriquecimiento semántico y validación de la calidad de los datos, mejorando la interoperabilidad y reutilización de los mismos.

Por último, destacar que resulta complejo ofrecer una solución genérica para todos los dominios ya que cada institución establece sus protocolos a la hora de publicar los datos como LOD con respecto a los vocabularios y repositorios utilizados. Además, las soluciones existentes en el campo de la identificación de fuentes para el enlazado de conjuntos de datos estructurados presentan ciertas carencias entre las que destacan la falta de validación de la calidad de los datos y la creación de contextos a partir de bases de datos estructuradas y colaborativas.

1.2 Objetivos y estructura de la tesis

El problema abordado en esta tesis es que a pesar de existir una importante cantidad de datos públicos y accesibles en Internet, no se garantiza que sean útiles para los usuarios que los consultan, no siempre se garantiza que sean confiables ni que puedan ser reutilizados de manera eficiente. En concreto aunque los datos sean abiertos y en formatos que permita su recolección y lectura no son fácilmente interpretables por los usuarios.

La hipótesis planteada en esta tesis para resolver este problema es la siguiente: *Aplicando un método de transformación a un conjunto de fuentes diversas de datos utilizando un modelo de enfoque multidimensional basado en RDF se pueden generar datos útiles y de calidad, enriquecidos y validados para los usuarios expertos y no expertos.*

El objetivo general de esta tesis es definir un modelo que a partir de un conjunto de datos obtenidos de diversas fuentes, facilite la publicación, enriquecimiento y validación de LOD generando información útil y de calidad orientada a usuarios expertos y no expertos.

Para la consecución del objetivo general se han propuesto los siguientes objetivos específicos:

1. Analizar los trabajos previos realizados sobre la integración de datos abiertos, así como los diferentes conceptos relativos a esta investigación: datos abiertos, calidad, uso de vocabularios y estándares y enriquecimiento semántico.

2. Definir e implementar procesos de recolección, integración y transformación de datos de fuentes diversas para obtener una única fuente de datos estructurados y marcados semánticamente..

3. Enriquecer semánticamente a través de KG y de otros *datasets* públicos.

4. Evaluar la calidad de los datos que se publican.

5. Generar una visualización orientado a usuarios no expertos a partir de un enfoque multidimensional.

6. Aplicar el modelo a varios casos de uso para analizar los resultados, refinar y generalizar el modelo.

El documento de esta tesis está estructurado en los siguientes apartados:

- Metodología de la investigación. El capítulo 2 describe la metodología seguida en este trabajo de tesis.

- Marco teórico y revisión bibliográfica. El capítulo 3 describe el análisis realizado de la literatura relacionada con los conceptos más destacados de esta investigación, como los datos abiertos, Web semántica, conjunto de datos multidimensionales y calidad en los datos. Además, se incluye la sección 3.6 que analiza otras aproximaciones para la publicación de LOD.

- Solución planteada. Definición de un modelo para la publicación de datos abiertos enlazados. El capítulo 4 describe el marco propuesto en este trabajo de tesis, en el se describen los seis pasos orientados a la publicación y explotación de LOD.

- Aplicación del modelo. Iteración 1. Escenario real de la red de abastecimiento de agua en la Comunidad Valenciana. El capítulo 5 describe la aplicación y evaluación de la primera versión del modelo planteado en esta tesis inicialmente compuesto por 4 pasos: (1) especificación de las fuentes de datos, (2) generación, (3) publicación y (4) explotación de los datos. Además en la sección 5.6 se describen los resultados obtenidos después de validar el modelo a través de un grupo de expertos.

- Aplicación del modelo. Iteración 2. escenario real *Open Data* Barcelona. El capítulo 6 presenta la segunda iteración del modelo en la que se describe la aplicación y evaluación del modelo refinado después de la primera iteración. Se describen los seis pasos que componen el modelo refinado: (1) mapeo de datos y preprocesamiento de fuentes, (2) modelado, (3) generación, (4) almacenamiento, (5) evaluación de la calidad y (6) explotación de datos. Además la sección 6.8 describe los resultados de la valoración mediante criterio de expertos sobre el modelo refinado.

- Valoración mediante criterio de usuarios. El capítulo 7 describe la validación realizada por un grupo de usuarios que validan los resultados obtenidos de la aplicación del modelo en la segunda iteración.

- Conclusiones. El capítulo 8 exponen las conclusiones obtenidas después del trabajo realizado en esta tesis doctoral. Además, se describen brevemente las líneas de trabajo futuro.

- Difusión de la investigación. El capítulo 9 se listan los trabajos publicados como fruto de esta investigación además de las diferentes actividades llevadas a cabo para la difusión.

2. Metodología de la investigación

El cambio siempre es dulce.

Aristóles

En este capítulo se presenta la metodología de la investigación que se ha utilizado en este trabajo de tesis. Se plantea la hipótesis a validar y el método aplicado para demostrar la validez de la misma.

2.1 Método de investigación aplicado

El problema abordado en este trabajo y la hipótesis para resolverlo están descritos al inicio de la sección 1.2. Para la validación de la hipótesis se plantea un experimento esperando que los resultados obtenidos la demuestren. Para el desarrollo de esta tesis se han seguido los siguientes pasos:

1. Análisis de la literatura en lo relativo a los conceptos más destacados de esta investigación como los datos abiertos, datos enlazados, conjunto de datos multidimensionales y calidad en los datos.

2. Estado de la cuestión en el que se analizan metodologías para la publicación de LOD con el objetivo de detectar puntos de mejora en las propuestas y trabajos previos.

3. Descripción del modelo. Esta fase describe en detalle las etapas que componen el modelo final propuesto para la demostración de la hipótesis.

4. Demostración. Aplicación y evaluación de la solución planteada en el paso anterior. Para realizar esta fase se plantea un método iterativo en el que al final de cada ciclo se obtiene una mejora del modelo propuesto.

   Para valorar e identificar posibles mejoras en el modelo planteado, se involucró en el proceso a quince expertos en datos abiertos enlazados. El método utilizado ha sido Delphi [Hsu and Sandford, 2007], técnica de recogida de información que permite obtener la opinión de un grupo de expertos a través de la consulta reiterada [Reguant Alvarez and Fonseca, 2016]. En la literatura se pueden encontrar trabajos que utilizan el método Delphi aplicado en diferentes campos como la tecnología, medicina, ciencias sociales, etc. De hecho se puede decir que este método es posiblemente uno de los más utilizados por los investigadores [Cabrero Almenara and Barroso Osuna, 2013, Hsu and Sandford, 2007].

   Se han planteado dos iteraciones. La literatura señala que suele ser suficiente dos o tres iteraciones para recabar la información necesaria; en este caso, con dos iteraciones se ha podido realizar la propuesta final del modelo. En cada iteración se recopiló la valoración de los expertos, así como los comentarios y sugerencias de mejora sobre la propuesta, mediante el uso de formularios. Se ha mantenido el anonimato entre ellos (solo conoce su identidad la autora de esta tesis).

   Como se puede observar en la figura 2.1 cada iteración está compuesta de 3 fases: (1) aplicación del modelo a un caso real, (2) valoración por parte de los expertos y (3) refinamiento del mode-

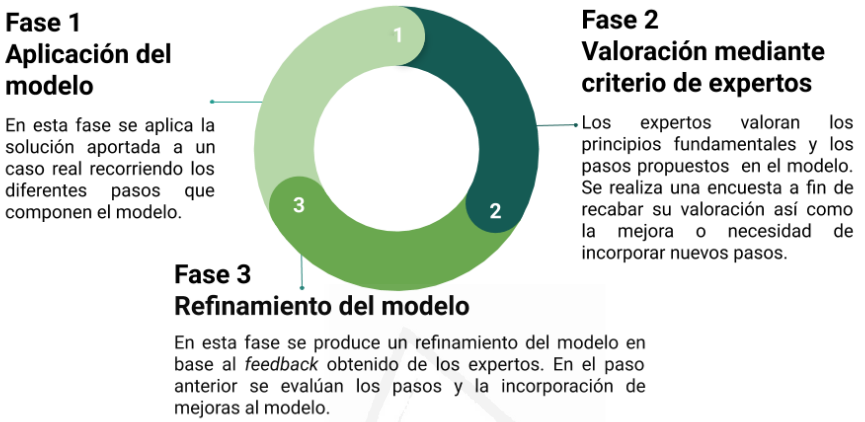


Figura 2.1: Ciclo de vida iterativo en el que cada iteración está compuesta por tres fases. Al final de cada ciclo se obtiene una mejora del modelo. Fuente: producción propia.

lo de acuerdo con los resultados de la fase de validación anterior. A continuación se describen brevemente las dos iteraciones realizadas. En cada una se aplica el modelo a un caso de estudio real y se valora mediante el criterio expertos.

Primera iteración:

- Se realiza una primera propuesta del modelo y se aplica a un caso real de un dominio concreto con una fuente de datos homogénea y estructurada. Los datos se han suministrado en formato *Excel Spreadsheet* (XLS).
- Se realiza la evaluación del modelo por parte de un grupo

de expertos y se recoger los resultados obtenidos de dicha evaluación.

- A partir de la retroalimentación recibida en la fase anterior, se mejora la propuesta inicial. De esta manera se refina el modelo con la valoración mediante el criterio de expertos.

Segunda iteración:

- Se aplica el modelo refinado a un caso real, de un dominio concreto de fuentes públicas de datos. En esta fase se recolectan datos de la plataforma de datos abiertos de Barcelona, *Open Data BCN*[1]. A diferencia de la primera iteración, aquí se introducen fuentes de datos heterogéneas. En el mejor de los casos los datos están estructurados y se suministran en formato *Comma-Separated Values* (CSV). Pero para la experimentación ha sido necesario incorporar datos extraídos de archivos *Portable Document Format* (PDF), de los cuales ha sido complejo extraer la información incluso con un *Optical Character Recognition* (OCR).
- Se realiza la evaluación del modelo mediante el mismo grupo de expertos que en la primera iteración.
- Después de la evaluación del modelo refinado y a partir de los resultados obtenidos, se comprueba que se han cumplido con las recomendaciones dadas en la primera iteración.

5. Validación. La validación del modelo propuesto se ha realizado, por un lado, mediante el criterio de un grupo de expertos, y por otro, mediante el criterio de usuarios basándose en los resultados obtenidos de la aplicación del modelo a un caso real.

[1] `https://opendata-ajuntament.barcelona.cat/`

- **Valoración mediante criterio de un grupo de expertos**. Se valora los principios fundamentales y los pasos que conforman el modelo propuesto mediante criterio de expertos. La selección del grupo de expertos no se ha realizado de manera aleatoria sino que se han seleccionado directamente, en función de su vinculación y experiencia profesional con la temática a evaluar, por tanto, se ha aplicado el método cuasi experimental, en el que una de sus características más destacable es que la selección de los miembros de un grupo de expertos no se realiza de manera aleatoria. Con ello se obtendría una validación cuasi experimental y sería la manera de formalizar la solución planteada en dos iteraciones.
- **Valoración mediante criterio de usuarios**. En este paso se realiza la valoración por usuarios de perfil no técnico. En esta punto se pretende obtener el grado de satisfacción de los usuarios con el resultado obtenido, quienes dan el criterio sobre si lo aplicado les resulta útil o no.

3. Estado de la cuestión

> Saber que sabemos lo que sabemos y saber que no sabemos lo que no sabemos, ese es el verdadero conocimiento
>
> *Nicolás Copérnico*

En este capítulo se ofrece una visión general de los conceptos e investigaciones anteriores relacionadas con los datos abiertos, cómo ha afectado en ellos la Web semántica, el estado actual de la reutilización de datos abiertos enlazados y los problemas que actualmente siguen vigentes. También se realiza una revisión bibliográfica sobre la publicación de LOD, metodologías y modelos existentes.

3.1 Open Data

"Los datos no son solo el petróleo del siglo 21 sino la tierra más fértil para que crezca nuestro conocimiento, creatividad e innovación".

David McCandless

Los datos son abiertos, si son accesibles y están disponibles con las características tanto técnicas como legales necesarias para que se pueden usar, reutilizar y redistribuir libremente para cualquier propósito y sin ningún tipo de restricción [The World Bank, 2013]. Hoy en día los

datos son un factor clave tanto para la innovación social como para el crecimiento económico (innovación, creación de empresas y eficiencia de las empresas), sin olvidar que tienen un tremendo valor comercial [European Data Portal, 2015].

Facilitar el acceso a los datos que se generan en las distintas instituciones a través de *Open Data*, abre un nuevo escenario de colaboración entre las diferentes instituciones y los ciudadanos, por ejemplo en lo que respecta a la reutilización y evaluación de esta información. Además, no solo los ciudadanos sino que las empresas, tanto las ya establecidas como los nuevos emprendedores, se han sumando a esta corriente utilizando *Open Data* para poder explorar mercados potenciales y desarrollar nuevos productos basados en los datos.

Dentro del contexto de las administraciones públicas, los datos abiertos facilitan la transparencia en los gobiernos, fomentan la participación ciudadana, ayudan a contribuir en la mejora de los servicios públicos y en una mayor eficiencia en la gestión de la información, ya que los datos abiertos se pueden obtener, reutilizar y redistribuir para compartirlos incluso con otras instituciones [Open Data Barometer, 2018]. Por ello, entre otras razones, muchas instituciones públicas y gubernamentales investigan diferentes formas de hacer visibles sus datos y publicarlos como LOD. Además, se trabaja en el desarrollo de nuevas interfaces que faciliten la interacción con los usuarios y puedan así tener acceso y hacer uso de los datos.

Sin embargo, para explotar todo el potencial de los datos abiertos y que realmente aporten valor (reutilizables y útiles para las instituciones y usuarios), se deben tener en cuenta diferentes aspectos clave como, la calidad, uso de estándares, facilidad en el acceso o que los formatos en los que se publiquen sean procesables automáticamente. Estas deberían ser condiciones indispensables para publicar un conjunto de datos que sea verdaderamente útil para los usuarios.

El uso de los datos abiertos en empresas, instituciones o por parte de los usuarios puede formar parte de procesos tan importantes como

la toma de decisiones o predicciones, por lo que en este sentido, un proceso de generación y publicación de LOD debe garantizar la calidad de sus datos (determina la confiabilidad).

Al fin y al cabo, todos los productores de datos deben estar comprometidos en la medida de sus posibilidades, con los principios de gestión de la calidad y desarrollar los mecanismos necesarios para mejorarla, reduciendo los errores o defectos que se producen en las diferentes etapas por las que debe pasar los datos hasta que se publican en un repositorio. Esto engloba desde las etapas más tempranas como la recolección y fiabilidad de las fuentes; normalización de los datos; integración de datos provenientes de fuentes heterogéneas; modelización y generación del RDF a partir de una ontología; enriquecimiento y publicación.

En 2015 se desarrollaron seis principios que sentarían las bases de cómo hacer accesibles, publicar y usar los datos [Open Data Charter (ODC), 2015]. Posteriormente, en 2018 se publicó el informe final para ayudar y guiar a las instituciones (gobiernos) en la publicación de datos abiertos, donde se ofrece orientación y ejemplos de buenas prácticas [Brandusescu and Lämmerhirt, 2018, Open Data Charter (ODC), 2018]. Estos son los seis principios:

1. Los datos deben estar, por defecto, en abierto. Las instituciones deben poner a disposición del público los datos pero a su vez han de garantizar que se cumple con las normativas sobre seguridad y protección de datos, identificando aquellos datos sensibles que deben continuar siendo privados. Esto supone un cambio muy significativo en lo que a la divulgación de datos se refiere y el impacto sobre la sociedad y la propia economía. Aportando gran valor para todas aquellas instituciones, empresas o usuarios que deseen reutilizar esos datos.

2. Deben estar actualizados, completos y exhaustivos. Se debe co-

nocer la fecha de publicación de los datos, tener acceso a otras versiones o publicaciones anteriores de los *datasets*. Se debe asegurar que los datos son relevantes, precisos y consistentes. Los datos tienen valor si son de calidad.

3. Accesibles y usables. Se debe garantizar el acceso a los datos, es decir que sean recuperables sin restricciones. Deben ser procesables por máquinas garantizado formatos que así lo permitan pero sin olvidar que las personas también deberían poder leerlos y entenderlos. Además, es necesario facilitar que puedan ser descubiertos.

4. Comparable e interoperable. Con la interoperabilidad se facilita el intercambio y reutilización de los datos y por ello es importante el uso de estándares para que los diferentes actores puedan entenderse. Además, los datos deben ser comparables y se deben poder establecer en diferentes ejes como el tiempo o localizaciones entre otros. De esta manera los datos se sitúan dentro de un contexto aumentando así su comprensión.

5. Para mejorar la gobernanza y la participación ciudadana. La transparencia que ofrecen las instituciones que ponen sus datos en abierto permite mayor interacción con los usuarios y otras instituciones. Además, facilita que los gobiernos rindan cuentas, ofreciendo una imagen de confianza y responsabilidad.

6. Para el desarrollo inclusivo y la innovación. Publicar los datos en abierto puede ayudar y estimular el desarrollo económico. Pero no solo hay que centrarse en publicar sino en promover su reutilización y explotación por parte de instituciones y usuarios.

La cantidad de datos que se generan crece rápidamente. Actualmente, muchas son las organizaciones públicas y privadas que están

generando grandes cantidades de datos, además, se suman diversas fuentes de datos, como los obtenidos de dispositivos inteligentes, sensores IoT o de las distintas redes sociales, que arrojan a Internet una cantidad ingente de datos cada minuto.

El estudio realizado a 18 iniciativas de datos abiertos en cinco ciudades, Barcelona, Chicago, Manchester, Amsterdam y Helsinki [Ojo et al., 2015], muestra la utilización y el impacto de los datos abiertos en estas ciudades. El estudio identifica siete dominios sobre los que impactan estas iniciativas: economía, educación, energía, medio ambiente, gobierno, turismo y transporte y como el propio estudio comenta, los dominios son similares a los contemplados en la literatura. Por ejemplo, se crean aplicaciones basadas en datos, servicios digitales y también revela un patrón de impacto inherente a la "economía de innovación abierta". En el estudio [Dong et al., 2017] se proporciona una explicación detallada de los conjuntos de datos para cada ciudad canadiense, incluidos los diferentes catálogos de datos y sus características. En ambos artículos se destaca la importancia de los datos abiertos y la innovación resultante en estas ciudades.

Las instituciones que trabajan con el patrimonio cultural también se unen a la filosofía de publicar sus datos en abierto. Las bibliotecas albergan colecciones en diferentes formatos que constituyen una parte esencial del patrimonio cultural de la sociedad. Gracias al desarrollo de internet y de las nuevas tecnologías, numerosas colecciones han sido publicadas en formato digital favoreciendo su acceso y ampliando su visibilidad. Además, muchas de ellas han comenzado a compartir sus metadatos como LOD para enriquecerlos y difundirlos. La *Library of Congress Linked Data Service*[1] proporciona acceso a los datos de autoridad, como por ejemplo *Library of Congress Subject Headings* o sobre áreas geográficas. En 2011, la Biblioteca Nacional Francesa pu-

[1] `http://id.loc.gov/`

blicó su repositorio de datos abiertos[2] agregando información sobre autores, obras y materias que provenían de diversas fuentes de datos (diferentes catálogos). La Biblioteca Nacional de España migró sus catálogos a RDF y se pueden consultar en `datos.bne.es` [Vila-Suero et al., 2013]. Del mismo modo ha ocurrido con el catálogo de la Biblioteca Virtual Miguel de Cervantes que se migró por primera vez a RDF en el año 2015 [Romero et al., 2018]. Recientemente, se ha creado la *International Community GLAM Labs*[3] que tiene como objetivo la reutilización de las colecciones de forma innovadora y creativa. Un *Lab* es un espacio físico o digital en que se ofrecen herramientas que explotan y reutilizan las colecciones digitales, así como *datasets* con licencias abiertas y ejemplos de utilización orientados a la investigación [S-C. et al., 2019]. Lideradas por la Biblioteca Británica,[4] numerosas instituciones han creado un *Lab* como por ejemplo la Biblioteca Nacional de España[5], la Biblioteca del Congreso de los Estados Unidos[6], la Biblioteca de Holanda[7] o la Biblioteca Virtual Miguel de Cervantes[8].

Como se observa los datos abiertos se expanden a muchos y diversos dominios. Según el último informe de 2019 sobre la madurez de los datos [European Data Portal, 2019], la primera categoría y dominio más popular son los datos abiertos sobre el sector público y el gobierno (ver figura 3.1).

Este hecho genera un creciente interés en incorporar estas enormes cantidades de datos externos (en muchos casos no estructurados) en las aplicaciones tradicionales dentro de la organización, en los propios procesos de la empresa e incluso en la toma de decisiones. Sin embar-

[2] `https://data.bnf.fr/`
[3] `https://glamlabs.io/`
[4] `https://www.bl.uk/projects/british-library-labs`
[5] `https://bnelab.bne.es/`
[6] `https://labs.loc.gov/`
[7] `https://lab.kb.nl/`
[8] `http://data.cervantesvirtual.com/blog/`

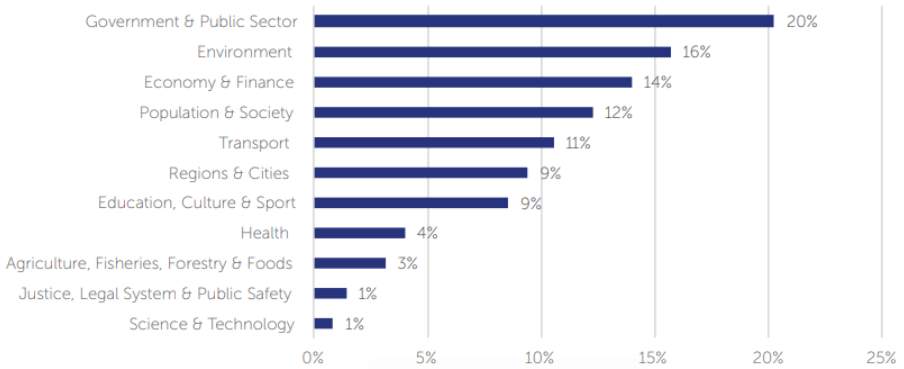


Figura 3.1: Dominios más populares para la publicación de datos abiertos según el último informe de 2019 sobre la madurez de los datos. Fuente: European Data Portal [2019].

go, esta oportunidad potencial que ofrece la información recopilada, a menudo no se aprovecha debido a que requieren técnicas avanzadas para el análisis de los datos, visualización y servicios que permitan la exploración de los mismos.

A este respecto, el uso de ontologías específicas, como por ejemplo en los dominios de ciudades inteligentes, como *Smart City Ontology* (SCO) [Bianchini et al., 2018, Rani et al., 2017], permite una exploración semántica de datos urbanos. Como se verá en la sección 3.2, *Linked Data* puede aportar la solución o al menos facilitar la explotación y reutilización de los repositorios de datos abiertos.

El informe de 2019 sobre la madurez de los datos abiertos en Europa [European Data Portal, 2019] (recopilación de datos disponibles en los portales de datos públicos en los países europeos), realiza una evaluación a partir de las dimensiones:

- Política, en lo que se refiere a políticas y estrategias para el fomento de los datos abiertos.
- Portal, mide funciones de acceso a los datos e interacción con los usuarios así como la cobertura y sostenibilidad del portal.
- Impacto, en el que se mide la reutilización y su impacto centrado en las áreas de política, social, ambiental y económica.
- Y por último, la dimensión de calidad en la que se mide la recolección, cumplimiento de estándares y calidad en el proceso de publicación de datos.

En sus conclusiones finales marca que el 9 % de los países (España, Francia y Alemania) cuenta con una política avanzada de datos abiertos, satisface las necesidades de usuarios avanzados y editores, y el nivel de calidad es muy bueno. El 25 % presenta buen nivel de madurez en todas las dimensiones, pero se observan problemas en la monitorización del impacto de los datos y su reutilización. El 44 % cuentan con una política de datos abiertos, pero presentan limitaciones en términos de publicación y reutilización de datos. Y el 22 % restante está en una etapa temprana de madurez para las cuatro dimensiones evaluadas.

Este informe hace especial énfasis en la calidad de los datos abiertos, su reutilización e impacto. Sin embargo en anteriores estudios se centraban en la preparación y madurez de los datos. Este cambio en los indicadores y dimensiones que se evalúan, marca claramente una nueva etapa en la publicación y reutilización de los datos.

La figura 3.2 muestra la clasificación de los países en la dimensión *portal*, la cual revela que los países con mayor madurez en esta dimensión son Francia (91 %), España (89 %), Irlanda y Chipre (ambos 86 %).

Un ejemplo de mejora producida con la apertura de los datos sería proporcionar información sobre el consumo y la disponibilidad de agua

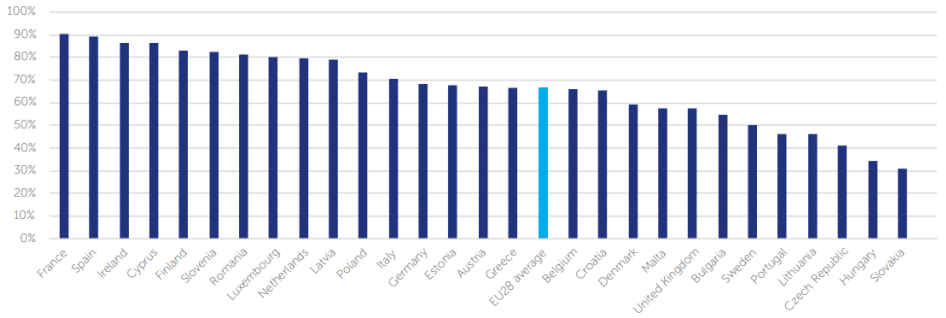


Figura 3.2: Grado de madurez de los países en la dimensión de *portal*. Fuente: European Data Portal [2019].

potable, así aumentaría la conciencia del usuario final y mejoraría la calidad en la toma de decisiones por parte de las empresas suministradoras [Curry et al., 2014]. Sería crucial mejorar el acceso a estos datos y fomentar el intercambio abierto de información, siguiendo con el ejemplo, sobre el suministro y consumo de agua, para minimizar el impacto e incluso llegar a resolver problemas existentes con los recursos hídricos [David et al., 2016], como ocurre en determinadas zonas de España en la que hay escasez de agua.

En cuanto a la organización y formato se refiere, los conjuntos de datos se organizan de manera diferente atendiendo al dominio en el cual se generan. Por ejemplo para las *smart cities*, habitualmente los *datasets* se organizan en categorías basadas principalmente en sus actividades gubernamentales, como la seguridad, cultura y ocio, medio ambiente, transporte e instalaciones de la ciudad o gastos [Portal, 2017]. En el dominio del patrimonio cultural y bibliotecas digitales publican diferentes conjuntos de datos, según temática, autores, épocas, etc.

El formato de un conjunto de datos abiertos hace referencia a cómo se estructuran y publican los datos tanto para humanos como para las máquinas. Deben publicarse sin procesar, correctamente estructurados y en formatos que faciliten la gestión y su reutilización. Se puede proporcionar una respuesta satisfactoria a las necesidades de los usuarios mediante el uso de formatos comunes, más sencillos como los ficheros CSV o Excel ambos muestran los datos estructurados pero el CSV es un formato abierto. También se pueden utilizar otros formatos más avanzados como RDF, *Extensible Markup Language* (XML) o *JavaScript Object Notation* (JSON) [Open Knowledge International, 2012].

Aunque los diferentes informes sitúan la publicación de datos abiertos en un grado de madurez aceptable a nivel europeo (34 % tiene grado alto) o el caso concreto de España (según `datos.gob.es`) el 80 % se sitúa en el nivel de 3 estrellas según la clasificación propuesta por Tim Berners-Lee, son formatos reutilizables, es decir, abiertos, estructurados, pueden ser procesados automáticamente, aún se pueden encontrar datos publicados en otros formatos no estructurados, propietarios y más complejos de procesar como PDF, Word, imágenes, audio o vídeo.

El *World Wide Web Consortium* (W3C) recomienda que los datos que se publiquen sean útiles y estén en formatos que faciliten el procesamiento automático y sean reutilizables. Por lo tanto, la preferencia desde la perspectiva de los datos, es que se publique en formatos abiertos y legibles por máquina.

Para que los datos estén disponibles, los editores normalmente organizan un catálogo central en el que se enumeran los diferentes conjuntos de datos. También se puede poner a disposición de los desarrolladores un API, que permita la gestión de estos datos para poder identificar y seleccionar aquellos necesarios en un momento dado y así evitar descargar todo el conjunto de datos. Normalmente estas APIs proporcionan diferentes métodos y criterios para tal fin. Otras iniciativas proponen un modelo conceptual que puede originar una ontología

de gobierno electrónico escalable y efectiva [Sourouni et al., 2010]. Además, se ha analizado la interconexión de Schema.org con vocabularios para mejorar el proceso de enriquecimiento de un conjunto de datos [Nogales et al., 2016].

En las ciudades existen sistemas complejos que generan grandes cantidades de datos, pero siguen existiendo desafíos relacionados con el estudio de técnicas visualización avanzadas y servicios que permitan la exploración más eficiente de esos datos. En este contexto, las ontologías descritas para las ciudades inteligentes, como SCO, proporcionan una herramienta eficaz para la exploración semántica de los datos urbanos [Bianchini et al., 2018, Rani et al., 2017].

En el análisis realizado en esta sección es donde se evidencia el problema planteado en esta tesis.

3.2 *Linked Data*

En 2001, Tim Berners-Lee expuso "La Web semántica no es una Web separada, sino una extensión de la actual, en la que la información tiene un significado bien definido, que permite que las computadoras y las personas trabajen mejor en cooperación"[9]. El objetivo de la web semántica es que cualquier entidad o relación entre entidades se pueda codificar en la web.

La gran cantidad de información que se genera y publica en Internet, junto con el gran número de fuentes de información y los problemas que se generan por la falta de semántica, provoca que en muchas ocasiones sea complejo acceder a la información que está disponible. Con la Web semántica, se ha buscado dotar de significado a los contenidos online, para que sea más sencillo compartir y reutilizar datos, y para que cuando los usuarios hagan una búsqueda, obtengan la información que realmente necesitan. No hay más que consultar los cam-

[9] `https://www.w3.org/standards/semanticweb`

bios recientes realizados por Google, en los que incorpora a su motor de búsqueda el algoritmo *Bidirectional Encoder Representations from Transformers* (BERT) [Devlin et al., 2018] , considerada la mayor actualización en los últimos años. Cómo ha cambiado el concepto de las búsquedas en las que ya no es tan importante la coincidencia de las palabras claves indicadas al realizar dicha búsqueda, sino el contexto y significado, la verdadera intención del usuario al hacer la búsqueda, en este caso a través de la Inteligencia Artificial.

Tim Berners-Lee [Tim Berners-Lee, 2006] propuso un nuevo modelo llamado *Linked Data* para publicar datos estructurados en la Web, basados en RDF [RDF Working Group, 2014], a fin de facilitar la conexión y reutilización entre ellos. *Linked Data*, como parte de la web semántica, se han convertido en la mejor opción disponible para la apertura de datos porque facilita la integración de estos en la Web [Bizer et al., 2009]. Tim Berners-Lee acuñó estos 4 principios básicos[10] necesarios para conseguir interconectar los datos:

1. Se deben utilizar URIs para identificar unívocamente los recursos publicados.
2. Se deben utilizar URIs sobre *Hypertext Transfer Protocol* (HTTP) para asegurar que se puedan buscar y localizar esos recursos.
3. Se debe suministrar información útil y fácilmente procesable sobre el recurso al que referencia el URI, haciendo uso de estándares como RDF o SPARQL.
4. Se deben incluir enlaces a otros repositorios y promocionar el descubrimiento de más información.

Para la comprensión de la Web semántica es necesario describir algunos términos esenciales.

[10] `https://www.w3.org/wiki/LinkedData`

RDF es el marco de descripción de recursos recomendado por el W3C para representar información en la web. Describe recursos en términos de tripletas y es indispensable para la estructura de los datos abiertos enlazados [McBride, 2004]. Una tripleta RDF está compuesta por sujeto, predicado y objeto, y describe la relación del sujeto con el objeto a través del predicado, enlazándolos. RDF puede representarse en diferentes formatos: Turtle[11], N-triples[12], RDF/XML[13], Notation3 (N3)[14], JSON-LD[15] y JSON[16]. Algunos como N-Triples, N3 y Turtle son formatos más sencillos de interpretar por las personas frente a RDF/XML cuya comprensión es más compleja. RDF/XML es la sintaxis normativa para escribir RDF según el W3C[17].

Resource Description Framework Schema (RDFS) es una extensión semántica de RDF que proporciona un vocabulario de modelado de datos que permite describir propiedades y clases de recursos RDF [Staab and Studer, 2013], ya que RDF no define los vocabularios que se utilizan en las declaraciones.

SPARQL [Harris et al., 2013] es el lenguaje de consulta W3C para repositorios RDF. Trabaja con grafos RDF [Prud et al., 2006] donde las fuentes de datos se identifican mediante URIs.

Web Ontology Language (OWL) es una extensión de RDF utilizada cuando la información contenida en los documentos debe ser procesada automáticamente por las aplicaciones. Permite describir el significado de la terminología usada en los documentos Web [World Wide Web Consortium (W3C), 2014, 2004].

Las ontologías ofrecen un modelo formal de conceptos de interés

[11] `https://www.w3.org/TeamSubmission/turtle/`
[12] `https://www.w3.org/TR/2014/REC-n-triples-20140225/`
[13] `https://www.w3.org/TR/REC-rdf-syntax/`
[14] `https://www.w3.org/TeamSubmission/n3/`
[15] `https://www.w3.org/TR/json-ld/`
[16] `https://www.json.org`
[17] `https://www.w3.org/TR/rdf-primer/#rdfxml`

(clases), características y atributos de cada concepto (propiedades) y restricciones de propiedad, que involucran un dominio de conocimiento específico en el mundo real [Noy et al., 2001, Guarino et al., 1998]. Las ontologías son una capa en el conjunto de estándares de W3C[18]. Una base de conocimiento se compone de una ontología y sus instancias (conjunto de clases y propiedades). Estas, proporcionan servicios para facilitar la interoperabilidad entre sistemas heterogéneos.

Los KG como DBpedia [Auer et al., 2007], Wikidata [Tanon et al., 2016] y YAGO [Rebele et al., 2016] son una rica fuente de información para enriquecer los conjuntos de datos originales. Cubren el conocimiento general, conocido como *cross domain*, en lugar del conocimiento relacionado con dominios específicos. Proporcionan información estructurada que permite las conexiones con otros repositorios externos (por ejemplo, GeoNames) y aspectos multilingües a través del acceso a descripciones y propiedades en diferentes idiomas.

Estas bases de conocimiento gratuitas, abiertas y basadas en LOD están aumentando su popularidad, promoviendo la publicación y reutilización de *Open Data*. Wikidata adopta un enfoque innovador al proporcionar un flujo de trabajo online para proponer la creación de nuevas propiedades que se discuten de manera colaborativa y si se llega a un consenso, la propiedad finalmente es creada por un administrador. En este sentido Wikidata[19] ha despertado gran interés al ofrecer nuevas oportunidades para la participación de la comunidad a fin de ahorrar tiempo y energía a los profesionales. Estar vinculados a Wikidata proporciona beneficios como:

- Mejores resultados en los motores de búsqueda, enriquecidos con la información proporcionada por los KG. Estar conectado a estos repositorios es clave para aumentar la visibilidad y establecer una fuerte presencia online.

[18] `https://www.w3.org/standards/semanticweb/`.
[19] `https://www.wikidata.org`

- Se abren nuevos caminos de validación entre diferentes recursos y hacia una mejor integración de los datos [Waagmeester et al., 2016]
- La experiencia se aporta a través de los voluntarios e investigadores a nivel mundial que pueden conectar los artículos con otras colecciones.
- Además, Wikidata permite la ejecución de consultas federadas SPARQL para conectar con una serie de bases de datos externas, incluidas Europeana, la Biblioteca Virtual Miguel de Cervantes y la Biblioteca Nacional de España [Wikidata, 2017].

Publicar como LOD proporciona las siguientes ventajas:

- Interoperabilidad con otros sistemas.
- Facilita compartir y enriquecer datos. Los datos están mejor organizados de manera que se facilita el acceso a ellos.
- Ofrece sistemas de búsqueda y navegación más robustos y fáciles de usar potenciados por la semántica.
- Facilita el descubrimiento de recursos en colecciones heterogéneas y distribuidas.
- Permitir que las máquinas entiendan los datos.
- Es una extensión de *World Wide Web*, no la reemplaza.

La publicación de grandes conjuntos de datos como datos abiertos enlazados es un desafío que requiere del diseño y la implementación de métodos personalizados para la transformación, gestión, consulta y enriquecimiento de los datos. Además, se debe garantizar la calidad de los datos. Existen diferentes métodos para evaluar la calidad del

conjunto de datos, algunos están orientados a ofrecer información sobre el grado de reutilización y apertura en sus datos. Se podría decir, en cierta medida, que cuanto más reutilizable es un conjunto, más valor tiene para los usuarios. Otras propuestas están más centradas en el análisis de dimensiones más profundas de los datos en sí, como la integridad, relevancia, grado de comprensión, completitud, validez o confiabilidad.

Tim Berners-Lee ofrece un sistema de clasificación para medir la calidad de los datos en función de la facilidad con la que pueden ser reutilizados por otros. Esta clasificación se conoce como el esquema de desarrollo 5 estrellas de LOD, promovido por W3C. Se establecen cinco niveles, de menos a más en función de facilidad de reutilización (ver figura 3.3). Para establecer el nivel en el que se clasifican los datos, tiene en cuenta el formato con el que se publican y cómo se presentan los datos. A continuación, se describe cada uno de los niveles que componen este esquema:

* OL de *Open License*. Se publican los datos en la Web bajo una licencia abierta pero en cualquier formato independientemente de lo difícil que sea procesarlos. Por ejemplo, documentos en formato PDF no legible o imágenes escaneadas, estos datos no son estructurados y son complejos de gestionar y reutilizar, es la forma más sencilla y primitiva de publicar datos.

** OL y RE del inglés *Open License y Machine-readable*. Se publican como datos estructurados pero el formato es propietario. Un ejemplo muy conocido es el formato XLS en lugar de una imagen de una tabla escaneada, este formato es propietario pero se puede procesar, exportar a otro formato y sigue siendo sencilla su publicación.

*** OL, RE y OF del inglés *Open License, Machine-readable y Open Format*. Se publican los datos estructurados utilizando formatos

no propietarios. Por ejemplo publicando los datos en un CSV o en OpenOffice, en vez de utilizar formato XLS. Se puede procesar sin necesidad de software propietario y sigue siendo sencillo de publicar.

**** OL, RE, OF y URI del inglés *Open License, Machine-readable, Open Format y Uniform Resource Identifier*. Se utilizan URIs para identificar recursos y propiedades de manera que se puedan referenciar fácilmente. Se pueden visualizar directamente en la web siguiendo estándares promovidos por W3C como RDF. Este nivel significa un salto cuantitativo en cuanto a la reutilización y publicación de datos abiertos pero requiere más esfuerzo y conocimientos técnicos (ayuda de expertos) para publicar los datos, así como más coste de mantenimiento.

***** OL, RE, OF, URI y LD del inglés *Open License, Machine-readable, Open Format, Uniform Resource Identifier y Linked data*. Este es el nivel más alto, los datos se enlazan con otras fuentes de datos lo que permite incrementar su valor. Al igual que pasaba en el nivel anterior, requiere más esfuerzo, conocimientos técnicos y un mayor coste de mantenimiento del conjunto de datos.

Existen otras técnicas que permiten evaluar la calidad de los datos. MELODA [Abella and Ortiz-de Urbina-Criado, 2014] es una métrica para evaluar el grado de reutilización de los datos abiertos. En su versión 3.10, evalúa en función de cuatro dimensiones: marco legal, estándares técnicos, accesibilidad de la información y modelo de datos. La valoración de una fuente de datos se basa en la medida de estas cuatro dimensiones. Cada dimensión tiene en cuenta de 3 a 5 niveles de madurez y cada nivel se pondera con un peso diferente (ver tabla 3.1). Para obtener la valoración de una organización que publica datos

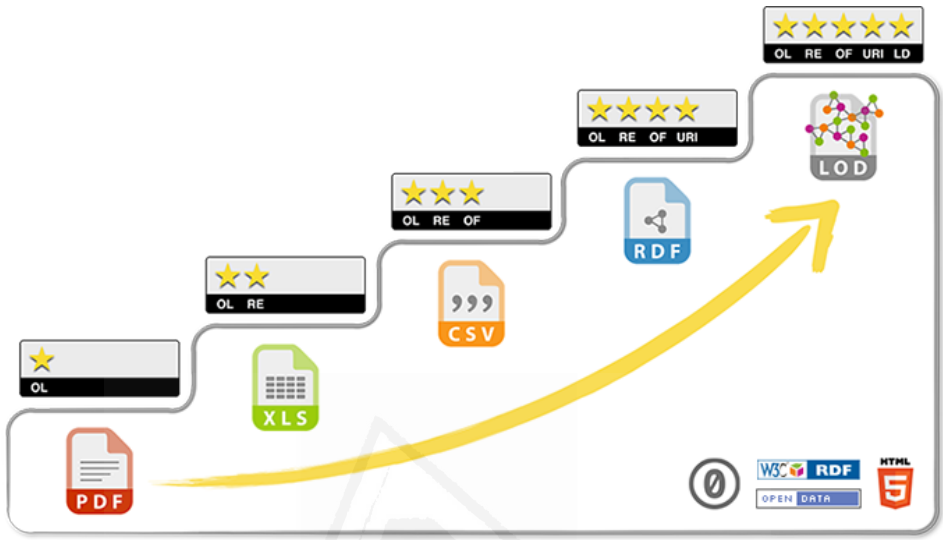


Figura 3.3: Sistema de clasificación 5 estrellas propuesto por Tim Berners-Lee. Fuente: `https://5stardata.info`

abiertos, se calcula la media de los valores obtenidos para cada una de las fuentes de datos publicadas.

FAIR (del inglés *Findable, Accessible, Interoperable and Reusable*) *Data maturity model* es un modelo basado en los principios de búsqueda, accesibilidad, interoperabilidad y reutilización [FAIR Data Maturity Model WG, 2020]. Establece un conjunto de indicadores y niveles de madurez para medir el grado de cumplimiento de estos cuatro principios, es decir el nivel de encontrabilidad, accesibilidad, interoperabilidad y posibilidad de reutilización de los datos. Además, proporciona pautas para mejorar el cumplimiento de estos principios [Wilkinson et al., 2016]. Inicialmente orientados a los datos de investigación, pueden aplicarse a cualquier dominio. Estos principios fueron

Tabla 3.1: Dimensiones y niveles para la métrica MELODA.

	Nivel	descripción	Peso
Marco Legal	1	Copyright	0,00 %
	2	Uso privado	25,00 %
	3	Reutilización no-comercial	50,00 %
	4	Reutilización comercial	75,00 %
	5	Reconocimiento solo	100,00 %
Estándares técnicos	1	Privativos	20,00 %
	2	Abiertos	60,00 %
	3	Estándar abierto con metadatos	100,00 %
Accesibilidad	1	Sin acceso	0,00 %
	2	Acceso web con registro	25,00 %
	3	Acceso directo vía web	50,00 %
	4	Acceso vía web con parámetros	75,00 %
	5	API o lenguaje de consulta	100,00 %
Modelo de datos	1	Sin modelo publicado	0,00 %
	2	Modelo con campos de datos	25,00 %
	3	Modelo con especificaciones de campos	50,00 %
	4	Modelo externo normalizado	75,00 %
	5	Modelo externo y generalizado	100,00 %

incorporados a los proyectos de *Horizon 2020* (Programa de Investigación e Innovación de la Unión Europea).

Otras iniciativas proporcionan criterios de calidad que permiten analizan los conjuntos de datos a partir de dimensiones como la precisión, integridad, licencia, relevancia y accesibilidad entre otras. En el estudio [Piscopo, 2016, Färber et al., 2018] se propone una lista de criterios de calidad de datos, para evaluar los KGs dentro del contexto de LOD. Este enfoque utiliza los conceptos de criterios, dimensiones y categorías originalmente propuestos en investigaciones previas sobre la calidad de los datos [Wang and Strong, 1996]. Estas iniciativas están más enfocadas en evaluar la calidad de los KG.

Se están haciendo muchos esfuerzos por publicar los datos como LOD y conforme se mejora en la madurez de los datos abiertos que se publican, es necesario atender otros aspectos clave como la calidad. Como se puede observar, existen diferentes técnicas para abordar la problemática existente con la calidad de los datos y la propuesta en este trabajo es evaluar la calidad de los datos de acuerdo con la propuesta de [Färber et al., 2018] pero sin perder de vista los principios FAIR y el esquema 5 estrellas propuesto por Tim Berners-Lee.

3.3 Las barreras en la reutilización de *Open Data*

Según el informe publicado por la Comisión Europea sobre la reutilización de *Open Data* [Portal, 2017], todavía quedan barreras tanto internas como externas que impiden a los usuarios la estandarización y automatización en la recolección y procesamiento de *Open Data*. El informe concluye con una serie de recomendaciones a seguir tanto el sector público como el privado.

En su informe actualizado [European Data Portal, 2019], se concluye que el 66 % de los países europeos presentan serios problemas en la publicación y reutilización de los datos.

En el estudio [Ruijer et al., 2017] sugieren que la interacción entre gobiernos, industrias y universidades podría abordar los problemas derivados de presupuestos ajustados o las limitaciones en la contratación de personal, lo cual dificulta la implementación de nuevos procesos, más inteligentes, haciendo uso de las nuevas tecnologías y explotando los resultados de la investigación.

En [Benitez-Paez et al., 2018], los usuarios citan que la falta de unas guías básicas para el uso y enriquecimiento de los datos disponibles, tiene un impacto negativo en el nivel de reutilización. En este informe

se sugiere la creación de un *kit* básico de reutilización que incluya una guía que los ayude a descargar, conectar, enriquecer y mostrar los datos publicados. De esta forma se podría ayudar a los usuarios a comprender cómo los diferentes conjuntos de datos abiertos podrían ser explotados de manera significativa.

Otros estudios [Link et al., 2017] muestran una serie de factores que podrían incidir directamente en el impacto de los datos abiertos, como la recolección mediante herramientas automatizadas, pero todavía presentan retos por resolver sobre todo en lo relativo a garantizar la privacidad, la calidad y el propio análisis de los datos.

Cuando se trata de considerar qué conjunto de datos usar, la calidad de los datos es un aspecto crucial y la falta de calidad o de su verificación desalienta a otros a contribuir y reutilizar los datos. Existen varias iniciativas para especificar y evaluar la calidad como ya se ha descrito en la sección 3.2 [Piscopo, 2016, Färber et al., 2018], MELODA [Abella and Ortiz-de Urbina-Criado, 2014] o FAIR *Data maturity model* [FAIR Data Maturity Model WG, 2020].

3.4 *Linked Data* y los datasets multidimensionales

El modelo entidad-relación (E-R) permite representar de manera simplificada los componentes y cómo se relacionan entre sí, para un determinado proceso. Es la normalización de la estructura de datos, donde se obtiene un diseño sin redundancia, se controla la integridad referencial y minimiza el espacio de almacenamiento. A pesar de ser el modelo más extendido no contribuye a facilitar las consultas. Una técnica mucho más potente para la interrogación de datos, es el modelo multidimensional.

En el contexto de los datos abiertos enlazados, los modelos multidi-

mensionales son la combinación de diferentes conjuntos de datos, que permiten la aplicación de técnicas de evaluación mediante estadísticas e indicadores [Hira and Deshpande, 2015, Carrasco et al., 2017]. Según el W3C, un conjunto de datos estadísticos comprende una colección de observaciones que se pueden organizar en un conjunto de dimensiones, atributos y medidas, conocidos como componentes [World Wide Web Consortium (W3C), 2017].

Los modelos de datos multidimensionales, pueden tener diferentes representaciones relacionales, incluido el esquema de estrella y el esquema *snowflake* [Kimball, 1996]. Ambos utilizan tablas de dimensiones para describir los datos agregados en una tabla de hechos. El más común es el esquema en estrella, cuya característica principal es que las tablas de dimensiones no están normalizadas. Esta representación parece una estrella (de ahí su nombre), en ella las tablas de dimensiones rodean la tabla central (la de hechos). Además, las tablas de dimensiones en este esquema, no se relacionan. La granularidad dentro de cada dimensión también está determinada por la necesidad de detalles. Por otro lado, el esquema *snowflake* refleja la organización jerárquica de las dimensiones y a diferencia del esquema en estrella, aquí las tablas de dimensiones están normalizadas y las jerarquías se representan en tablas separadas, lo que permite mejor comprensión de los niveles de clasificación definidos en la dimensión.

El vocabulario RDF Data Cube [Cyganiak and Reynolds, 2014] es una recomendación del W3C para la publicación de datos multidimensionales (como estadísticas) en la Web. Define específicamente las dimensiones, atributos y medidas utilizadas en el conjunto de datos y se basa en los vocabularios RDF existentes:

- *Simple Knowledge Organization System* (SKOS) es una iniciativa del W3C para representar la estructura básica y esquemas conceptuales como los encabezamientos de materia, tesauros o taxonomías.

- *The Statistical Core Vocabulary* (SCOVO), es un vocabulario básico para describir información estadísticas. Está basado en tres conceptos: *dataset*, *item* y *dimension*.

- *Dublin Core Terms*[20] mantenido por *Dublin Core Metadata Initiative* (DCMI) para la descripción de los metadatos en la que están incluidos los 15 elementos básicos, más, las propiedades, clases, tipos de datos y esquemas de codificación del vocabulario. Cada término se identifica con su URI y se representan en RDF para que puedan ser utilizados en LOD.

- *Friend of a Friend* (FOAF) para describir personas, sus relaciones y actividades[21].

- ORG[22], como vocabulario para la descripción de estructuras organizativas que tiene en cuenta una amplia variedad de ellas.

Además, este vocabulario es compatible con *Statistical Data and Metadata Exchange* (SDMX)[23], estándar ISO que proporciona un metamodelo para la descripción de datos en cualquier dominio estadístico, facilitando el intercambio y compartición de datos y metadatos estadísticos entre organizaciones.

En la literatura se pueden encontrar algunos ejemplos de uso de modelos multidimensionales y *Linked Data* [Kalampokis et al., 2013, Carrasco et al., 2017]. En [Etcheverry and Vaisman, 2012, Etcheverry et al., 2015], se propone un nuevo vocabulario pero, como una extensión del vocabulario RDF Data Cube que admite operaciones OLAP avanzadas, como *rollup*, *slice*, *dice* y *drill-across*, utilizando consultas estándar SPARQL.

[20] `https://dublincore.org/specifications/dublin-core/dcmi-terms/`
[21] `http://xmlns.com/foaf/spec/`
[22] `https://www.w3.org/ns/org#`
[23] `https://sdmx.org/`

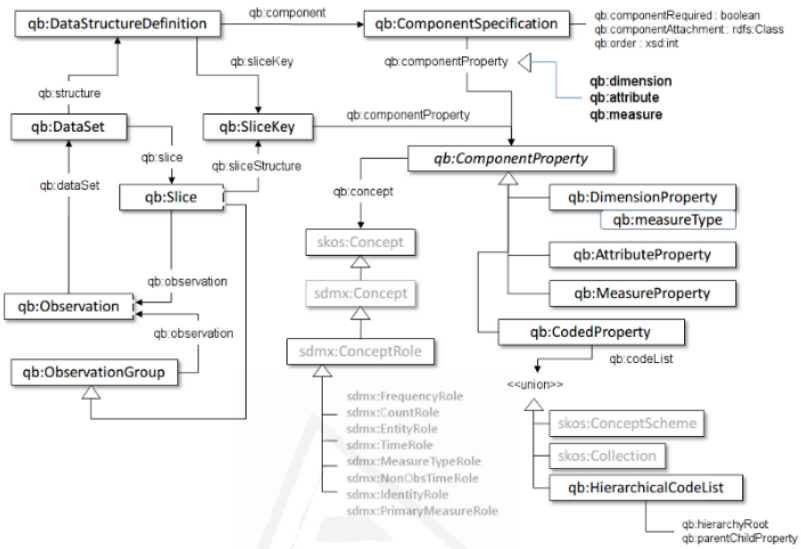


Figura 3.4: Resumen de los términos clave y su relación en el vocabulario RDF Data Cube. Fuente: W3C *The RDF Data Cube Vocabulary*.

Varios institutos nacionales de estadística, incluyendo Italia[24], Irlanda[25], Grecia[26], Escocia[27], UK[28] y Japón [Matsuda et al., 2018] proporcionan sus datos estadísticos como LOD basados en el vocabulario RDF Data Cube.

En [Klímek et al., 2018], la publicación de estadísticas oficiales sobre pensiones como LOD y basado en el vocabulario SKOS y RDF Data Cube, ilustra cómo se reutilizan los datos en las aplicaciones y

[24] `http://datiopen.istat.it/index.php?language=eng`
[25] `http://data.cso.ie/sparql`
[26] `http://linked-statistics.gr/sparql`
[27] `https://statistics.gov.scot/`
[28] `http://statistics.data.gov.uk/sparql`

cómo contribuyen a los indicadores estadísticos en combinación con otros LOD. AirBase, es el conjunto de datos sobre la calidad del aire en Europa, mantenido por la Agencia Europea de Medio Ambiente. Representa en un cubo de datos RDF, la información sobre la contaminación del aire , además, se ha vinculado a las bases de conocimiento de YAGO y DBpedia [Galárraga et al., 2017].

En paralelo, emergen nuevas aplicaciones para la visualización y exploración estadística de los datos basados en el vocabulario RDF Data Cube. Las infraestructuras de *Tecnología de la Información y Comunicación* (TIC) a menudo tienen muchas restricciones para integrar nuevas aplicaciones. Las aplicaciones tradicionales del tipo cliente-servidor dependen de la disponibilidad del lado del servidor. Por el contrario, existen aplicaciones del lado del cliente como CubeViz.js [Abicht et al., 2017]. CubeViz.js es una aplicación *lightweight* basada en el vocabulario RDF Data Cube que permite la visualización y exploración de los datos a través de conexiones a un punto de acceso SPARQL o a un conjunto de datos RDF (publicados en formato N-Triples, Turtle, N-Quads[29] o TriG[30]). En [Folmer et al., 2019] se presentan cuatro métodos para la visualización de datos e identifican casos de usos potenciales de un conjunto de datos, por ejemplo visualizaciones 3D o la integración de LOD en herramientas de *Business Intelligence*.

Sin embargo, siguen existiendo desafíos relacionados con el modelado de cubos de datos como LOD y con los diferentes enfoques para abordarlos [Kalampokis et al., 2015]. Por ejemplo, cómo definir el número de medidas, cómo modelar los datos espacio-temporales [Mijović et al., 2016], o la falta de vocabularios especializados [Varga et al., 2016]. Actualmente, las instituciones desarrollan sus propias soluciones ad-hoc dificultando tanto la publicación como la posterior reutilización

[29] `https://www.w3.org/TR/n-quads/`
[30] `https://www.w3.org/TR/trig/`

de estos datos. En consecuencia, resulta complejo integrar diferentes fuentes y desarrollar herramientas de software genéricas [Kalampokis et al., 2015].

El gran valor de los modelos multidimensionales es poder combinar diferentes conjuntos de datos y explotarlos en su conjunto. Y el paradigma de los datos abiertos enlazados puede ayudar a conseguirlo. Es importante en este aspecto, el uso de estándares confiables, promovidos por organizaciones como W3C para la publicación de datos, ya que facilitan su reutilización. Y al respecto, cabe mencionar que el W3C recomienda el vocabulario RDF Data Cube [Cyganiak and Reynolds, 2014] para la publicación de datos multidimensionales.

3.5 Almacenamiento de RDF

Como se ha comentado anteriormente el rápido desarrollo de la Web semántica ha provocado un aumento de datos RDF en la Web. En consecuencia, han emergido diferentes técnicas y sistemas para almacenar datos en RDF. El almacenamiento eficiente de RDF es algo discutido con anterioridad en la literatura [Pan et al., 2018, Faye et al., 2012]. Existen varias formas de almacenar datos RDF comúnmente conocidos como almacenes de tripletas. Estos admiten mecanismos de almacenamiento de datos, inferencia, actualización, escalabilidad, distribución y un punto SPARQL, entre otros. Recientemente han aparecido diferentes opciones basadas en Javascript [World Wide Web Consortium (W3C), 2018]. Sin embargo, otros enfoques utilizan directamente el conjunto de datos final, evitando así requisitos técnicos y, en ocasiones, complejos como la instalación y la configuración del servidor donde alojar el repositorio. También se puede almacenar en `datahub.io` que es un gran repositorio de almacén de datos que después puede consultarse para los casos en los que no es posible disponer de una infraestructura para almacenar el RDF.

Algunos de los sistemas de almacenamiento RDF existentes que proporcionan un punto de acceso SPARQL son los siguientes:

- Eclipse RDF4J[31] conocido anteriormente como OpenRDF Sesame, es un *framework* de código abierto en Java para el procesamiento y gestión de datos RDF. Incluye la creación, análisis, almacenamiento, razonamiento y consultas con RDF y datos enlazados. Proporciona un API para la conexión con otros sistemas de almacenamiento en RDF y un punto de acceso SPARQL. Además, RDF4J admite los formatos RDF más utilizados como RDF/XML, Turtle, N-Triples, N-Quads o JSON-LD. Dispone de una interfaz web para su gestión. Está disponible a través de Apache Maven[32] o mediante SDK[33].

- Virtuoso Universal Server[34], es una plataforma de almacenamiento híbrida que combina la funcionalidad de un sistema tradicional de base de datos relacional con XML y RDF. Proporciona tanto la opción de trabajar con una base de datos relacional como un punto de acceso SPARQL. Tiene versión de código abierto conocida como OpenLink Virtuoso[35] y la comercial.

- Stardog[36] sistema de almacenamiento en RDF que ofrece la validación de restricciones de integridad en los datos, indexación por lotes, motor de inferencia y aprendizaje automático entre otros. Además dispone de integración con Apache Lucene[37] para

[31] `http://rdf4j.org/`

[32] `https://maven.apache.org/`

[33] `https://rdf4j.org/download/`

[34] `http://virtuoso.openlinksw.com`

[35] `http://vos.openlinksw.com/owiki/wiki/VOS`

[36] `https://www.stardog.com/`

[37] `https://www.stardog.com/docs/man/query-search`. Lucene es una librería de código abierto que permite integrar funciones de indexación y búsquedas a texto completo `https://lucene.apache.org/`

realizar búsquedas a texto completo. Es de uso comercial pero dispone de una licencia gratuita de un año para fines académicos.

- 4store[38] sistema de almacenamiento RDF eficiente, escalable y estable. Ofrece un cliente web para la gestión. El código está disponible bajo licencia GNU *General Public License*.

- En D2RQ[39] es un sistema para acceder a bases de datos relacionales como si fueran grafos RDF virtuales (de solo lectura) sin necesidad de disponer de un almacén de datos RDF. Es de código abierto[40] y se publica bajo la licencia *Apache Version 2.0*.

3.6 Metodologías para la publicación de *Linked Open Data*

A medida que se publican más datos abiertos en la Web también evolucionan las prácticas y directrices para la publicación de LOD. En [Lee, 2017] se propone un flujo de trabajo para el ciclo de vida de datos vinculados basado en cuatro componentes: (1) adquisición, (2) método de aprendizaje de ontología, (3) almacenamiento RDF, y (4) un sistema de análisis. En [Lnenicka and Komarkova, 2019], realizan un análisis e identifican las limitaciones e inconvenientes de los marcos actuales y proponen una metodología para publicar LOD con el uso de la computación en la nube.

El W3C *Government Linked Data Working Group* propone una guía para ayudar en el acceso y reutilización de *Open Government Data* [Working Group Note, 2014]. Además de esta guía, existen otras publicaciones que proponen modelos de ciclo de vida que comparten

[38] `https://github.com/4store/4store`
[39] `http://d2rq.org/`
[40] `https://github.com/d2rq/d2rq`

fases comunes como, especificar, modelar y publicar datos utilizando formatos abiertos. Por ejemplo, [Bernadette Hyland, 2011] proporciona un ciclo de vida compuesto de siete fases: (1) identificar datos, (2) modelar, (3) nombrar con URIs, (4) describir, (5) convertir a RDF, (6) publicar y (7) mantener.

En el trabajo realizado [Villazón-Terrazas et al., 2011], los autores proponen un conjunto preliminar de pautas metodológicas para ayudar en la generación, publicación y explotación de los datos vinculados del gobierno. Su ciclo de vida consiste en cinco fases: (1) especificar, (2) modelar, (3) generar, (4) publicar y (5) explotar. Todas ellas capturan las tareas o fases que requiere un flujo de trabajo de la gestión tradicional de la información, pero proporcionan límites diferentes entre estas fases [Working Group Note, 2014].

Por otro lado, en junio de 2018 se publicó el informe final [Open data charter, 2018], en el que se propone una serie de guías para facilitar tanto a la administración como a las empresas asociadas, publicar y compartir datos para un uso específico y en áreas temáticas concretas. Las investigaciones preliminares se centraron en áreas temáticas concretas, y se planteaban cómo usar los datos abiertos para ayudar en la lucha contra el cambio climático, la corrupción y también aplicarlos en la agricultura. El marco propuesto en este estudio se compone de cinco fases: (1) producción de datos, (2) compartir, (3) procesar, (4) acciones, (5) mecanismo de respuesta.

Después de la revisión del estado de la cuestión se identifican características comunes entre los distintos marcos orientados a la publicación y explotación de datos abiertos vinculados. Sin embargo, se observa algunas carencias de aspectos clave a tener en cuenta en la publicación de LOD, puede ser que no se usen o bien han sido omitidas. Por ejemplo, como el uso de repositorios externos como base de conocimiento para el enriquecimiento de los datos (por ejemplo usando Wikidata o GeoNames), y la inclusión de un paso para realizar la

validación de LOD evaluando la calidad de los datos que se publican.

4. Solución planteada. Definición de un modelo para la publicación de datos abiertos enlazados

> Si buscas resultados distintos
> no hagas siempre lo mismo.
>
> *Albert Einstein*

Este capítulo describe el modelo propuesto en este trabajo de tesis. Este ha sido refinado en dos iteraciones (ver capítulo de metodología de la investigación 2) y este apartado describe el modelo final refinado después de las dos iteraciones.

El modelo propuesto está basado en el ciclo de vida de [Villazón-Terrazas et al., 2011] que incluye las principales pautas metodológicas orientadas a la publicación y explotación de LOD. El enfoque propuesto en este trabajo mejora el proceso original planteado por Villazón ya que incluye un nuevo paso para evaluar la calidad de LOD. Hoy en día la generación y gestión de datos de buena calidad debe ser un requisito previo para la explotación y posterior intercambio de datos. Los usuarios deben poder explotar los datos abiertos para que tengan un impacto. Este nuevo paso de evaluación de LOD está basado en la metodología propuesta por [Färber et al., 2018] y se ha adaptado a las especificidades de los repositorios de cubos de datos. Además, se han enriquecido los datos originales mediante conexiones a repositorios externos. Y finalmente, se facilita la explotación del repositorio

mediante un punto de acceso SPARQL público orientado a usuarios expertos y un cuadro de mando orientado a usuarios no expertos, a fin de facilitarles la interacción con los datos.

4.1 Diseño del modelo

El modelo ha sido descrito con *Business Process Model and Notation* (BPMN), notación gráfica estandarizada, sencilla y clara para el modelado de procesos de negocios [1] (ver figura 4.1). En él se describen los pasos que conforman el proceso para la publicación y explotación del conjunto de datos: (1) especificación de las fuentes datos y preprocesamiento, (2) modelado, (3) generación, (4) publicación, (5) evaluación de la calidad y (6) explotación. Cada paso está representado por una actividad que puede ser implementada por una tarea (unidad mínima de trabajo dentro de un proceso) o por un subproceso (compuesto por más de una tarea o subprocesos). En cada paso se implementan una o más acciones que deben realizarse para completar esa actividad (invocar un servicio, acceder a las fuentes de datos, generar el RDF, cargar los datos en un repositorio, etc.). Además, se proporcionan algunas herramientas y técnicas para ayudar en su implementación.

El método describe claramente un proceso, por tanto, se ha optado por *Business Process Management* (BPM) para su correcta definición. Existen enfoques [Pérez et al., 2007] que consideran BPM como una parte fundamental de los modelos *Computation Independent Model* (CIM) y BPMN resulta adecuado para usarse en este nivel. Además, relacionan el paradigma BPM con la Arquitectura Dirigida por Modelos del inglés *Model-Driven Architecture* (MDA)[2].

[1] `https://www.omg.org/spec/BPMN/2.0.2/PDF/`
[2] `https://www.omg.org/mda/`

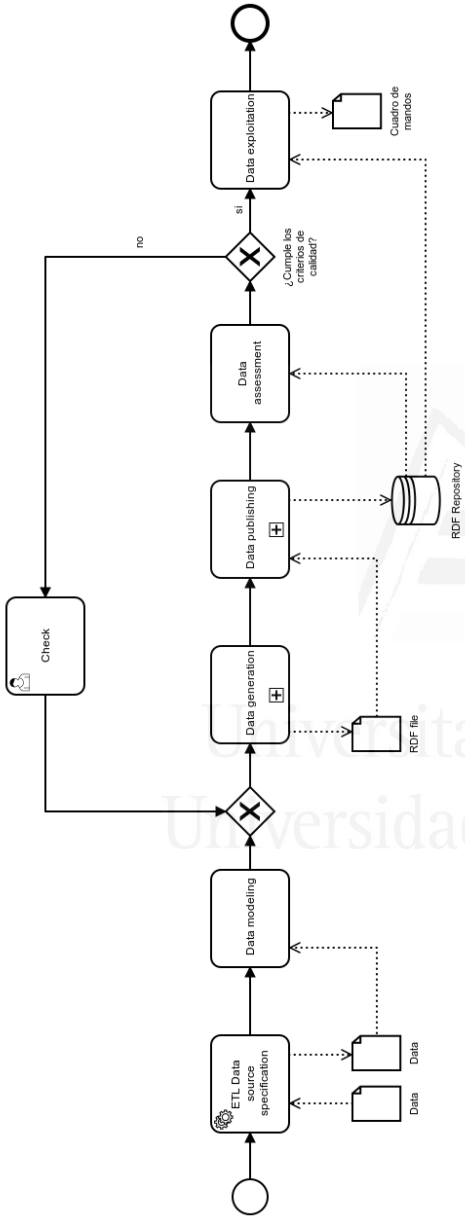


Figura 4.1: Modelado en BPMN de la propuesta para la publicación de LOD. Fuente: producción propia.

Este modelo se basa en los siguientes principios fundamentales:

- Interoperabilidad. La interoperabilidad es un principio clave para facilitar el intercambio y la reutilización de los datos, por tanto es importante el uso de estándares y proporcionar metadatos legibles por máquina. Es una de las máximas que debe cumplir un conjunto de datos.

- Reutilización. Este principio hace referencia al uso de los datos por parte de terceros. Los datos abiertos enlazados son un recurso abierto y deben ser accesibles por cualquier persona u organización que quiera consultarlos por lo que es necesario facilitar el acceso y la explotación del conjunto de datos. Su reutilización aporta una serie de beneficios como la creación de nuevos productos y servicios basados en los datos, estimulando el desarrollo económico y de conocimiento.

- Enriquecimiento semántico como la interconexión entre los datos que se generan y las bases de conocimiento. Permite la extensión del modelo de datos con la inclusión de nuevas propiedades mediante la ejecución de consultas federadas SPARQL.

- Calidad. La calidad en los datos no solo hace referencia a la ausencia de errores sino que es un aspecto mucho más amplio que abarca características como la accesibilidad, integridad, consistencia, interoperabilidad, etc. Los datos publicados deben ser útiles y confiables para los usuarios, incentivando así su reutilización.

En las siguientes secciones se describe cada uno de los seis pasos que componen este modelo.

4.2 Especificación de las fuentes de datos

Como se ha comentado previamente, las instituciones generan y manejan enormes cantidades de datos que están en distintos formatos, vocabularios y se almacenan en distintas bases de datos o repositorios. Trabajar con fuentes de datos heterogéneas dificulta el acceso y reutilización de estos datos. Esto requiere de un procesamiento previo para que puedan ser útiles, realizando tareas de limpieza, verificación, filtrado y transformación de datos.

Atendiendo a esta necesidad, el primer paso del modelo es la integración de los datos procedentes de diferentes fuentes de datos para que puedan ser utilizados posteriormente. Esta fase aborda los problemas derivados de la falta de semántica, uso de diferentes formatos, vocabularios e incluso repositorios externos utilizados por las instituciones, lo cual provoca falta de homogeneidad en los datos. En este paso se realizan las siguientes acciones: (1) selección de las fuentes (según la información que se necesite), conexión y recolección; (2) análisis y preproceso (preparación y normalización); y (3) generación de una vista integrada.

Para esta tarea de integración de datos (en ocasiones con datos complejos) se presenta un enfoque basado en el diseño de procesos *Extract, Transform and Load* (ETL) que facilitan el procesamiento y la gestión de los datos. Existen diversas herramientas con las que se pueden diseñar y ejecutar procesos ETL y su elección dependerá de las necesidades de la institución, por ejemplo en función del volumen de datos, conexiones con los almacenes de datos, capacidad de orquestación y automatización de los procesos, integración con otros servicios o el tipo de licencia de software entre otras.

Algunas de las herramientas disponibles para el diseño de ETL son

Data Integration de Pentaho (Kettle)[3]. Pentaho es un software integral que ofrece una plataforma para la integración de datos, orquestación y desarrollo de cuadros de mandos. Data Integration es de código abierto y forma parte de esta plataforma y facilita el acceso, la preparación y el análisis de datos no estructurados, con el objetivo de normalizar los datos obtenidos de fuentes de datos heterogéneas. Otra solución, pero comercial, es Microsoft Suite BI con su servicio *SQL Server Integration Services (SSIS)* para la integración y migración de datos.

Primero se realiza una selección de las fuentes en función del propósito. Existen diferentes sitios donde consultar conjuntos de datos abiertos disponibles como *Data Portals*[4], *The Linked Open Data Cloud*[5] o el Portal Europeo de Datos[6].

Una vez seleccionadas las fuentes se crean las entradas de datos para la conexión y recolección de los datos según el formato. El formato de un (*dataset*) se refiere a cómo se estructuran y publican los datos para humanos y máquinas. Elegir el formato correcto mejora la gestión y la reutilización. Si bien, el formato más común utilizado por las organizaciones para publicar sus datos es CSV, por ser un formato abierto, sencillo de entender, muy reutilizable, legible y fácilmente procesable por una máquina, existen propuestas más avanzadas que utilizan formatos como XML, RDF y JSON, que proporcionan un mayor nivel de información en lo que a la semántica se refiere [Open Knowledge International, 2012]. También se debe tener en cuenta que en algunos casos, los datos estadísticos se suelen representar utilizando XLS, formato propietario más comprensible y legible dentro de este

[3] `https://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+%28Kettle%29+Tutorial`

[4] `http://datacatalogs.org/` Listado de portales de datos abiertos

[5] `https://lod-cloud.net/` Listado de conjuntos de datos publicados como LOD. Actualmente contiene 1.255 conjuntos de datos con 16.174 enlaces (mayo de 2020)

[6] `https://www.europeandataportal.eu/data/datasets`

dominio, pero tanto sus macros como fórmulas complican su gestión.

Tras la selección de la fuentes, una vez están los datos en el sistema, estos se analizan y procesan (preparación y normalización). Sería recomendable recopilar la información sobre el conjunto de datos, su modelo de datos, propósito o las relaciones. Debido a la variedad de formatos, vocabularios y repositorios externos, aparecen problemas muy comunes, como errores en el texto, errores tipográficos, abreviaturas, diferentes idiomas, falta de información, semántica e incluso el hecho de tener que afrontar una desambiguación como en el caso de localizaciones [Candela et al., 2019]. En consecuencia, es necesario realizar un preprocesamiento que incluya un conjunto de analizadores realizando implementaciones propias o usando herramientas de extracción, transformación y carga [Bansal, 2014] para normalizar la información contenida en los datos fuente. Y por último se genera una vista integrada de los datos recolectados y tratados que se utilizarán el los pasos siguientes.

Es importante tener en cuenta que, aunque las fuentes de datos pueden diferir entre instituciones, esta propuesta pretende ser genérica para facilitar su aplicación a cualquier dominio. Este proceso requiere de la identificación de puntos comunes en las fuentes de datos que permita una correcta unión. Una vez que las fuentes de datos originales se tratan como un todo, se requieren varias tareas adicionales, como limpiar y normalizar los datos. Como resultado de este proceso semiautomático, se devuelve un archivo único con la información integrada que finalmente se utiliza para crear el RDF en un fase posterior.

4.3 Modelado de datos RDF

Una vez se han seleccionado las fuentes y preprocesados los datos recolectados para crear un vista integrada, el siguiente paso es la transformación de los datos a un modelo multidimensional, que incluye como

componentes: dimensiones, medidas y atributos. En este paso se realizan las siguientes acciones: (1) selección de la ontología adecuada para el modelado del dominio, (2) en caso de no tener una adecuada se creará priorizando la reutilización de recursos y vocabularios existentes, y (3) diseño del URI. La dimensión representa las perspectivas o entidades respecto a las cuales se quiere mantener los datos organizados proporcionando información sobre las observaciones, por ejemplo, el tiempo o la localización en la que se produce la observación. Los componentes de una medida representan el hecho observado (por ejemplo, la población de una región o la renta per cápita). Los componentes de un atributo sirven para calificar las observaciones, por ejemplo, la especificación de las unidades de medida o metadatos adicionales, como el estado de la observación (por ejemplo, oculto o verificado).

Es necesario seleccionar la ontología con la que modelar el dominio de los datos. La recomendaciones para la publicación de LOD propuesta por el W3C *Government Linked Data Working Group* (Grupo de Trabajo de Datos Vinculados del Gobierno del W3C) recomienda, siempre en la medida de lo posible, la reutilización de vocabularios estandarizados para facilitar la inclusión y expansión de la Web [Working Group Note, 2014]. Sin embargo, en algunos casos, es necesario diseñar una nueva ontología que permita la modelización del dominio en concreto. Esta nueva ontología se puede crear reutilizando otras existentes (la recomendación es reutilizar siempre que se pueda) o bien partiendo desde cero. Existen diferentes herramientas que facilitan esta tarea, por ejemplo, Protégé[7] [Musen, 2015].

En este aspecto, la recomendación del W3C para la publicación de datos multidimensionales es el uso del vocabulario RDF Data Cube [Cyganiak and Reynolds, 2014]. Utilizar este vocabulario aporta

[7] *Framework* de código abierto para el desarrollo de ontologías y otros sistemas inteligentes `https://protege.stanford.edu/`

beneficios importantes. En primer lugar, el conjunto de datos se publica en un formato legible por máquina y no propietario en lugar de proporcionar archivos estáticos como CSV y PDF. Además, las observaciones individuales son direccionables, es decir, permite el uso de datos de terceros mediante la creación de referencias. Y por último, hay que tener en cuenta que existen otros componentes y herramientas basados en RDF Data Cube, como CubeViz.js (aplicación ligera para la visualización de datos multidimensionales) [Abicht et al., 2017], lo que facilita su reutilización.

Siguiendo las pautas de diseño para la publicación de *Linked Data* [Tim Berners-Lee, 2006], los recursos deben identificarse mediante un URI. Su diseño debe ser simple, estable y fácil de gestionar. Existen algunos estudios que proponen guías para un diseño eficiente [Sauermann et al., 2008, Villazón-Terrazas et al., 2011], por ejemplo, utilizar URIs autodescriptivos, definir la estructura base (`base_URI`), usar barras o *hash* (#) para separar los diferentes fragmentos del URI, o definir el URI para el conjuntos de datos (`base_URI/dataset`) entre otras.

4.4 Generación de datos

Una vez se han seleccionado las fuentes, preprocesado los datos y seleccionada la ontología, el siguiente paso es la generación de los datos en RDF y este paso incluye la definición de un método para dicha transformación. En este paso se realizan las siguientes acciones: (1) generación de los datos en RDF según la ontología seleccionada en el paso anterior (sección 4.3), y (2) enriquecimiento semántico.

Para generar el conjunto de datos en RDF se pueden utilizar diferentes herramientas que permitan definir las transformaciones de los datos a cómo los representa la ontología, es decir en este punto se debe cumplir que las instancias generadas se ajusten a la ontología selec-

cionada en el paso anterior. Para realizar la transformación a RDF se puede implementar un script o una aplicación, usar RML-Mapper[8] o bien OpenRefine. El proceso de conversión a RDF puede realizarse en un proceso por lotes o de forma más interactiva (por ejemplo, mediante aplicaciones gráficas).

En este contexto es necesario resaltar dos herramientas muy representativas en los trabajos de web semántica:

1. Jena[9] es un *framework* de trabajo de código abierto en Java para trabajar en web semántica. Permite la definición y manipulación de grafos RDF. El grafo se representa como un "modelo" abstracto en el que las clases se utilizan para representar recursos, propiedades y literales.

2. OpenRefine[10] es una aplicación de código abierto que facilita la transformación de los datos en bruto a un formato legible por máquina. Las transformaciones o acciones a realizar las define el usuario y se almacenan en un proyecto. Posteriormente, se realiza un mapeo (de forma gráfica) del proyecto a un esqueleto RDF y su posterior exportación en un formato RDF. El esqueleto de alineación del esquema RDF especifica cómo se generarán los datos RDF a partir de los datos de origen. Las celdas en cada registro de datos se colocarán en nodos dentro del esqueleto.

Para cumplir con el requisito para la clasificación más alta del esquema 5 estrellas y con el cuarto principio sobre la publicación de datos abiertos enlazados, es necesario enlazar con otras fuentes de datos. En la transformación de los datos a RDF se incluye la fase de enriquecimiento y conexión con otros repositorios. Los conjuntos de datos son mucho más ricos, útiles y reutilizables cuando están vinculados con

[8] `https://github.com/RMLio/RML-Mapper`

[9] `https://jena.apache.org/documentation/rdf/index.html`

[10] `https://github.com/OpenRefine/OpenRefine`

otros repositorios. Y en este aspecto, RDF facilita la interoperabilidad y la definición de dichos enlaces.

Estos enlaces que permiten vincular diferentes repositorios se describen mediante la relación *owl:sameAs* y contribuyen a crear una mayor conectividad, y mayor riqueza promovida por LOD.

En general, el proceso de enriquecimiento se realiza en dos pasos:

1. Análisis de la información textual para descubrir posibles enlaces a recursos externos. En ocasiones estos recursos externos (fuentes de datos candidatas para el enlazado), no están disponibles en un formato adecuado para el proceso automático y no siguen los estándares para la Web semántica. En este caso, y siempre que sean datos abiertos, se puede recolectar dicha información y publicarla como LOD, pero esto supone crear el repositorio LOD y no reutilizar uno existente.

2. Validación manual de los enlaces candidatos realizada por expertos.

En la actualidad existen diferentes herramientas que pueden ayudar a disminuir la complejidad de las tareas, por ejemplo *Mix'n'match*[11] que permite listar entradas de bases de datos externas, realiza una comparación y ofrece una correspondencia con entradas de Wikidata de una manera rápida y simple.

Para el enriquecimiento de los datos cabe mencionar algunos de los repositorios más representativos y utilizados como fuentes de datos externas:

- GeoNames, base de datos geográfica disponible y accesible a través de varios servicios web (suministra un API de consulta de información). Además permite el enlace de información textual

[11] `https://tools.wmflabs.org/mix-n-match/`

a ubicaciones geográficas. Actualmente es uno de los repositorios externos más utilizados [Acheson et al., 2017].

- DBpedia es un proyecto cuyo objetivo es extraer contenido estructurado de la información creada en varios proyectos de Wikimedia. Es un grafo de conocimiento que almacena los datos en un formato legible por máquina.

- Wikidata es una base de conocimiento de edición colaborativa organizada por la Fundación Wikimedia y también es un KG. Es una base de datos orientada a documentos, enfocada en elementos que representan temas, conceptos y objetos.

4.5 Publicación de datos

Una vez generado el RDF el siguiente paso es publicar que consiste en: (1) la creación del repositorio y carga de los datos RDF, (2) inclusión de los metadatos descriptivos asociados al conjunto de datos, y (3) hacerlo accesible para que pueda ser reutilizado. Al menos debe incluir un URI para su descarga. Este paso es fundamental y se puede considerar el eje central del proceso, de hecho le da su nombre al modelo.

Como se ha visto en la sección 3.5 existen diferentes opciones para almacenar los datos RDF, y dependiendo de los servicios que se necesiten y recursos de los que se disponga, marcarán una elección u otra. Algunos de ellos admiten mecanismos de almacenamiento de datos, inferencia, actualización, escalabilidad, distribución y un punto de acceso SPARQL. En este punto es aconsejable validar y verificar la integridad, para este propósito existen diferentes técnicas y herramientas que ayudan a mejorar la exactitud y consistencia de los datos. Por ejemplo, a través de las restricciones, *Shapes Constraint Language*

(SHACL[12]) para RDF4J[13]. También existen herramientas de depuración de datos dirigidas por pruebas (test-driven data-debugging) que se basan en la generación de consultas SPARQL de prueba, por ejemplo RDFUnit[14] o la propuesta [Kontokostas et al., 2014].

Una vez cargados los datos RDF en el repositorio, es importante utilizar vocabularios para describir y facilitar su descubrimiento. *Vocabulary of Interlinked Datasets* (VoID) [World Wide Web Consortium (W3C), 2011] facilita esta tarea. Es un vocabulario RDFS para describir el conjunto de datos RDF y, como contempla su especificación[15], cubre cuatro áreas: metadatos de carácter general (basados en *Dublin Core*), metadatos de acceso (protocolos de acceso al conjunto de datos), metadatos estructurales (estructura y esquema) y, además, incluye metadatos para describir los enlaces entre *datasets* y su vinculación. Proporcionando información sobre la licencia de acceso y explotación de los datos se facilita que los usuarios conozcan las condiciones y los términos de uso. En general, esta información se especifica en RDF mediante relaciones como `dcterms:license` y `dcterms:rights`, ya sea integrado en el conjunto de datos o en un archivo VoID separado.

Para mejorar la visibilidad y descubrimiento del conjunto de datos se pueden publicar el sitemap (generados por ejemplo a través del punto de acceso SPARQL) y también dar de alta el nuevo conjunto de datos generado en diferentes portales como *The Linked Open Data Cloud*[16] o en *European Data Portal*[17].

Dado que la publicación directa del conjunto de datos final reduce las tareas de mantenimiento (en ocasiones complejas), se propone

[12] `https://www.w3.org/TR/shacl/`
[13] `https://github.com/eclipse/rdf4j/`
[14] `http://aksw.org/Projects/RDFUnit.html`
[15] `https://www.w3.org/TR/void/`
[16] `https://lod-cloud.net/`
[17] `https://www.europeandataportal.eu/en/about/be-harvested-us`

como mínimo publicar el RDF como un archivo que sea accesible por otros, por ejemplo utilizando plataformas como DataHub.io en el que se incluyan los metadatos con la información de licencia descritos mediante VoID.

El conjunto de datos debe cumplir con las 5 estrellas del esquema de clasificación propuesto por Tim Berners-Lee y con los principios de *Linked Data* descritos en la sección 3.2. Además, hay dos aspectos clave que deben cumplir:

- Reutilización, para ello se deben haber creado con formato ontológico, vocabularios y deben estar identificados mediante URIs. En este aspecto, si el conjunto de datos viene descrito utilizando vocabularios como VoID, facilita su descubrimiento y reutilización (VoID incluye metadatos para describir la estructura y esquema del conjunto de datos lo que facilita su consulta e integración de los datos).

- Interoperabilidad, que posibilita crear enlaces semánticamente equivalentes mediante herramientas que lo permitan.

4.6 Evaluación de LOD

En este paso se evalúa la calidad de los datos antes de que sean públicos y puedan ser reutilizados. Esta validación está basada en la propuesta descrita en [Färber et al., 2018]. En ella se propone un conjunto de criterios para evaluar la calidad de los datos en los KGs en el contexto LOD. Este enfoque utiliza los conceptos de categoría, dimensión y criterio originalmente propuestos en investigaciones previas sobre la calidad de los datos [Wang and Strong, 1996].

Un criterio de calidad de datos es una función con valores comprendidos entre 0 y 1 que da valor a una característica en particular como la disponibilidad o la precisión. Los criterios se agrupan en una

dimensión que a su vez se agrupan en categorías. La tabla 4.1 muestra las categorías y dimensiones utilizadas para la evaluación de la calidad.

Tabla 4.1: Listado de las dimensiones a medir para la evaluación de la calidad de los datos, agrupadas por categoría.

Categoría	Dimensión
Categoría intrínseca	Precisión, Fiabilidad, Consistencia
Categoría contextual	Relevancia, Integridad, Actualidad
Categoría representacional	Facilidad de entendimiento, Interoperabilidad
Categoría de accesibilidad	Accesibilidad, Licencia, Enlazado

Los criterios propuestos por [Färber et al., 2018] para evaluar la calidad de los datos se han adaptado a las especificaciones de los repositorios de cubos de datos.

4.7 Explotación de los datos

Este paso cubre la explotación del conjunto de datos como resultado del proceso de transformación. La publicación de datos como LOD permite su reutilización y el uso de vocabularios estándar basados en RDF mejora la interoperabilidad, la reutilización y la explotación por parte de otras instituciones. En este paso se plantea la explotación de un repositorio LOD según el público al que se orienta, dando solución tanto a usuarios expertos como no expertos:

1. Para intentar explotar todo su potencial, es necesario proporcionar un cuadro de mando que permita a los usuarios no expertos y sin demasiados conocimientos sobre las TIC, interactuar con el conjunto de datos. Para realizar esta tarea se propone el uso de CubeViz.js[18], es una aplicación en JavaScript *standalone* para

[18] `http://cubeviz.aksw.org/`

el descubrimiento y visualización de datos estadísticos. Genera un *widget* de navegación facetado que permite filtrar de manera interactiva las observaciones que se visualizan en los gráficos.

2. Se habilita un punto de acceso público SPARQL para facilitar el acceso y la reutilización del conjunto de datos por usuarios expertos y que además permita la conexión automática de aplicaciones. Un punto de acceso SPARQL no solo facilita el acceso a los datos, sino que también permite las consultas federadas y que estas se ejecuten en otros interfaces con acceso SPARQL.

Por último, destacar que los datos enlazados mejoran la inferencia de nuevos conocimientos al descubrir nuevas relaciones y analizar automáticamente el contenido de los datos, como por ejemplo, identificar posibles inconsistencias [W3C]. Se han realizado varios experimentos en este área [Ren et al., 2017, Colucci et al., 2017].

5. Aplicación del modelo. Iteración 1

> La inteligencia consiste no solo en el conocimiento, sino también en la destreza de aplicar los conocimientos en la práctica.
>
> *Aristóteles*

En este capítulo se describe la aplicación y evaluación de la primera versión del modelo descrito en la sección 4. En esta primera iteración de la fase de demostración de la metodología aplicada, se proponen 4 pasos a seguir para publicar LOD: (1) especificación de las fuentes de datos (mapeo y preprocesamiento), (2) modelado y generación del RDF, (3) publicación, y (4) explotación. Posteriormente a la aplicación del modelo al caso real y siguiendo la metodología propuesta para esta investigación, se realiza una valoración mediante el criterio de expertos para determinar si la aplicación y fases planteadas son correctas o necesitan alguna modificación.

5.1 Introducción

La figura 5.1 muestra la propuesta inicial del modelo compuesto por cuatro fases y adaptado al caso de uso en particular. Para esta aplicación se han utilizando los datos de una empresa de suministro de agua

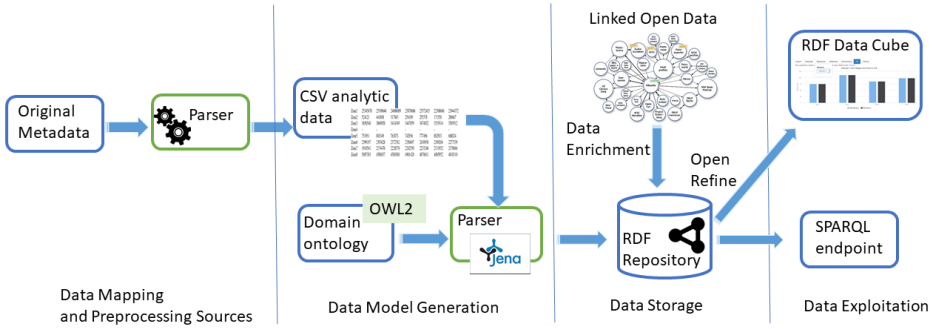


Figura 5.1: Propuesta inicial del modelo aplicado al caso real de la empresa de suministro de agua para la publicación de LOD.

de la Comunidad Valenciana que gestiona todos los procesos relacionados con el ciclo completo del agua: captación, tratamientos de agua para el consumo, transporte y distribución.

Esta zona del Mediterráneo que se ha utilizado como escenario de un caso real, ha sido y sigue siendo una zona castigada por la falta de agua dulce y teniendo en cuenta las previsiones internacionales, estos problemas se agravan. Según los estudios, existen diversos factores que afectan directamente y con alto impacto al problema del agua en esta región, entre esos factores se pueden citar: el crecimiento de la población y urbanización de zonas, turismo e industrialización, globalización y el cambio climático, provocando una disminución en la frecuencia de las precipitaciones y el consiguiente aumento de la sequía. Esta creciente escasez de agua y las incertidumbres derivadas del cambio climático, refuerzan la necesidad de adaptar tanto las políticas de agua como las políticas de planificación de tierras que impactan en la gestión del agua [Thivet and Fernandez, 2012].

Los datos utilizados en este caso real corresponden al período com-

prendido entre 2008 y 2014. Con respecto al consumo de agua en las ciudades de la Comunidad Valenciana, se suministraron alrededor de 476(hm^3) en 2014. Aproximadamente, el 25 % de este volumen (94 hm^3) era agua no registrada, bien por problemas en la red de suministro (fugas, interrupciones y fallos de red) o por problemas a la hora de la medición (errores y fraudes de clientes). El 75 % restante ($382hm^3$), corresponde al agua registrada, es decir, agua que ha sido medida [INE, 2018]. Además, contrastando estos datos con el consumo de agua en los hogares españoles durante los años analizados (2008-2014), es importante destacar que la Comunidad Valenciana, aunque no es la región española más poblada, sí que es la región que tuvo mayor consumo medio de agua (164,43 litros por habitante y día) durante el período estudiado.

Teniendo en cuenta estos datos, se aplica el modelo (compuesto de 4 pasos) al caso real de la empresa de suministro de agua con el objetivo de enriquecer y publicar los datos como LOD, de manera que estos se puedan explotar y sean útiles. Se busca facilitar el análisis y posterior toma de decisiones sobre el impacto de las diferentes medidas y datos disponibles (fugas, fallos de suministro, número de habitantes) y su relación con el consumo de agua, con el fin de reducirlo. Todo ello ayudaría a la gestión sostenible de los recursos hídricos naturales.

En las siguientes secciones se describe la aplicación del modelo a este escenario real, detallando cada uno de los pasos y las conclusiones extraídas de la experimentación.

5.2 Especificación de las fuentes de datos

En esta sección se describe como se ha realizado la especificación de las fuentes de datos para este caso real, siguiendo las pautas dadas

en la descripción del modelo (sección 4.2): (1) selección de las fuentes, conexión y recolección; (2) análisis y preproceso (preparación y normalización); y (3) generación de una vista integrada.

En este primer paso se lleva a cabo el mapeo y preprocesamiento de los datos, es decir se realiza la recolección, análisis y normalización mediante el uso de ETL y analizadores para obtener información normalizada de los datos de origen. Como se ha comentado en el capítulo 3, las instituciones utilizan diferentes formatos para publicar sus datos como CSV, XLS, incluso en formatos como XML, RDF o JSON que proporcionan más información sobre los datos de origen. Debido a esta heterogeneidad de las fuentes (tanto en lo relativo al formato como vocabularios) es necesario este paso de preprocesamiento.

Selección de las fuentes, conexión y recolección

En este caso real se suministran los ficheros fuente a procesar. El formato de partida de los datos originales es XLS y se sitúa en las dos estrellas de la clasificación propuesta por Tim Berners-Lee (esquema de desarrollo 5 estrellas del LOD promovido por W3C) en la que describe que los datos se publican como datos estructurados, pero el formato es propietario. La mayoría de los datos se han anonimizado por razones de privacidad.

Análisis y preproceso

Los datos originales suministrados (ver ejemplo de la figura 5.2) describen zonas y subzonas hidrográficas, valores de las diferentes mediciones realizadas en el tiempo del agua suministrada, registrada, renovaciones de la red de abastecimiento, acometidas, fugas en la red de transporte o en la distribución, escapes y roturas (en este caso solo se registra las roturas detectadas de manera visual). Además, en los

datos se detectaron indicadores como fugas de agua o roturas en la red de abastecimiento.

Para la integración de los datos se ha realizado un proceso ETL diseñado mediante la herramienta Data Integration de Pentaho (Kettle). Este proceso carga los datos de los diferentes archivos facilitados por la empresa (por ejemplo figura 5.2) y los preprocesa. En general, no se encontraron grandes complicaciones, sin embargo, surgieron algunos inconvenientes por la falta de semántica e información sobre la estructura de los datos, como por ejemplo en las medidas en las que algunos datos resultaron ser porcentajes y no valores absolutos. Además, se detectaron campos que carecían de valor debido a valores desconocidos, fallos en la medición o bien por errores producidos en la generación de los ficheros. Estos campos se muestran con el carácter "-" como se puede observar en el fichero original suministrado para esta experimentación (figura 5.2).

Generación de una vista integrada

La salida de este paso es un CSV, archivo de datos estructurados legible por máquina, y formato no propietario. La tabla 5.1 muestra un extracto de los datos obtenidos y que por razones de privacidad han sido anonimizados. Se listan un conjunto de zonas en las que se han realizado mediciones sobre el agua suministrada a lo largo de un periodo de tiempo. Por ejemplo en la zona 1.4 no hay registro de consumo y no hay detalle de si se ha producido un error o si realmente no ha habido suministro de agua. Los datos están preparados para que puedan ser modelados en el paso siguiente.

Tabla 5.1: Listado de consumo de agua en cada uno de las zonas en un periodo de tiempo determinado.

Zone	measure1	measure2	measure3	measure4	measure5	measure6	measure7
Zone1.1	25345678	25190646	24006089	23878686	23572415	22308686	21944372
Zone1.2	524121	441808	317843	254109	255378	171530	206667
Zone1.3	3656540	3669858	3414349	3447059	3474832	3519314	3565912
Zone1.4	-	-	-	-	-	-	-
Zone1.5	751951	803349	763073	742936	777496	802933	688024
Zone1.6	2399197	2393628	2372762	2356697	2430938	2309204	2277539
Zone1.7	1910541	2174476	2228578	2242550	2233246	2131932	2178066
Zone1.8	5087383	4588837	4580388	4901428	4870611	4885952	4818310

199	186	167	183	289	418	427	469	493	420
0	0	0	0	0	0	0	0	0	0
123	134	119	95	118	109	146	105	109	90
-	0	-	-	-	0	-	0	-	-
18	21	16	9	42	14	35	31	24	18
107	71	65	51	85	76	69	81	61	51
25	14	15	11	17	14	7	16	18	22
39	38	43	22	420	161	153	151	195	123
25	42	39	43	15	23	36	61	52	30
32	37	32	31	117	100	76	115	71	80
23	35	18	21	-	7	53	55	23	47
23	30	28	26	37	38	42	45	25	26
83	102	104	87	68	105	119	104	83	102
42	37	56	42	122	232	188	192	161	109
4	9	7	5	9	7	11	11	9	10
120	97	144	115	237	248	252	250	277	177
234	228	201	141	82	280	358	258	212	107
157	135	199	147	150	113	145	137	135	105
25	22	14	17	42	31	39	23	20	11
18	2	8	8	30	22	38	26	20	19
35	35	25	20	29	38	35	41	41	19
18	31	20	20	57	64	78	60	50	52

Figura 5.2: Ejemplo de fichero fuente suministrado por la empresa de suministro de agua para la aplicación del modelo.

5.3 Modelado de datos RDF

En este paso se realiza la generación del modelo de datos siguiendo las pautas dadas en la definición del modelo: (1) selección de una ontología adecuada para el modelado del dominio (en el caso de no encontrar una adecuada crearla reutilizando vocabularios existentes, siempre que sea posible), (2) diseño del URI, (3) generación del RDF, y (4) enriquecimiento semántico.

Selección de la ontología

Uno de los principales objetivos en esta primera experimentación es capturar, consolidar e integrar datos de diferentes fuentes relacionadas con el escenario del caso real (datos sobre el agua). Inicialmente, se ha optado por diseñar un enfoque semántico para el intercambio y reconciliación de los datos, mediante el cual se utiliza un modelo ontológico para la integración de datos. Razón por la cual se desarrolla una nueva ontología (en OWL) para describir los indicadores utilizados por la empresa de suministro de agua (ver anexo A). La ontología representa los datos de observación del agua junto con los metadatos descriptivos correspondientes, incluida la información sobre el tipo y unidad de los datos, los metadatos de procedencia, la ubicación de la observación y el momento en que se observan los datos.

Diseño del URI

La tabla 5.2 muestra cómo se define los URIs para la identificación de los recursos (define la ruta a los recursos utilizando descripciones explícitas de las entidades). Los puntos suspensivos representan la estructura base del URI (el prefijo común del URI es `http://www.khaos.uma.es/ontologies/lucentia/ontologies/water.owl#`) y los

Tabla 5.2: Patrones para las entidades definidas en el dominio de la empresa de suministro de agua.

Entidad	Patrón
Zone	.../zone/*
Water supply	.../hasNumberOfWaterSupplies/*
Water leaks	.../hasLeaksDistributionNetwork/*
Water not recorded	.../hasWaterNotRecorded/*
Length supply network	.../hasLengthSupplyNetwork/*

asteriscos un valor particular. Los prefijos y espacios de nombres utilizados en el conjunto de datos se listan en la tabla 5.3.

Generación del RDF

Para realizar el mapeo entre los datos obtenidos del paso anterior y la ontología se ha implementado un parser en Java utilizando Apache Jena. El proceso de generación del RDF se muestra en pseudocódigo 1 en el cual se destacan las tareas principales en la generación y enlazado del RDF.

En este paso se genera RDF en formato N-Triple, formato de texto plano, sencillo que facilita la generación de ficheros de gran tamaño, maximizando la interoperabilidad entre sistemas. Como se puede ver en el listado 5.5 la tripleta viene dada por la secuencia sujeto, predicado y objeto separados por un espacio en blanco. Las tripletas están separadas por un . al final. Jena facilita la creación de un fichero con el grafo RDF (un grafo en Jena se denomina modelo y se representa por `Model`) en diferentes formatos (NTRIPLES, JSONLD, RDF/XML, N3, RDF/XML Plain, TURTLE o RDF/JSON), la parte de código encargada de crear el fichero de salida se muestra en el listado 5.3.

Pseudocódigo 1 Procedimiento para la generación de RDF para el modelo de datos del agua.

Input: CSV datos del agua suministrado
Output: Formato RDF para los datos del agua modelados con la ontología propuesta

1: **procedure** GENERARDF(e)
2: Crea modelo
3: Define espacios de nombres
4: Genera URI
5: Carga fichero CSV con los datos de subZones
6: Crea RDF Subzones
7: Genera relaciones subZones y Wikipedia
8: Genera relaciones subZones y GeoNames
9: **return** RDF del modelo

```
public void writerRDF(Model model, String fileName) throws
    IOException{
        FileWriter out = new FileWriter(fileName + ".nt" );
        try {
            model.write( out, "N-TRIPLES" );
        }
        finally {
            try {
                out.close();
            }
            catch (IOException closeException) {
                // ignore
            }
        }
    }
```

Figura 5.3: Ejemplo de código Jena (Jena writer N-TRIPLES) para escribir datos RDF en formato N-Triple.

Tabla 5.3: Prefijos y espacios de nombres utilizados en el conjunto de datos.

Prefijo	URI
void	`http://www.w3.org/TR/void#`
rdf	`http://www.w3.org/1999/02/22-rdf-syntax-ns#`
owl	`http://www.w3.org/2002/07/owl#`
rdfs	`http://www.w3.org/2000/01/rdf-schema#`
qb	`http://purl.org/linked-data/cube#`
water	`http://www.khaos.uma.es/ontologies/lucentia/ontologies/water.owl#`
wd	`http://www.wikidata.org/entity/`
gn	`http://www.geonames.org/ontology#`

Enriquecimiento semántico

Para facilitar la reutilización y la interoperabilidad se han definido enlaces a recursos externos. Para realizar esta tarea se deben seleccionar los recursos adecuados con los que vincular los datos. En este sentido, los repositorios de LOD sobre la Comunidad Valenciana, en cierto modo, están limitados, pero existen algunas iniciativas como `datos.ign.es`[1] que proporcionan información geográfica como LOD. Además, en este sentido, se pueden utilizar repositorios *cross domain* y conjuntos de datos geográficos internacionales para enlazar las zonas y subzonas de los datos suministrados (las zonas y subzonas corresponden a localizaciones reales que se han anonimizado). El listado 5.4 muestra un ejemplo de enlazado con Wikidata y GeoNames utilizando la relación *owl:sameAs*. Wikidata actúa como un vínculo proporcionando enlaces a otros conjuntos de datos como GeoNames. Gracias al enriquecimiento, se pueden usar propiedades adicionales para proporcionar información contextual en las consultas SPARQL que no estaban disponibles en la fuente de datos original como, la población, coordenadas latitud longitud o el área administrativa.

[1] `http://datos.ign.es/` es una iniciativa del Instituto Geográfico Nacional (IGN) para proporcionar información semántica de sus recursos.

```
water:zone1
owl:sameAs
<http://www.wikidata.org/entity/Q935589> .
water:zone1
owl:sameAs
<http://www.geonames.org/2522419/agost.html> .
```

Figura 5.4: Ejemplo de enlazado con Wikidata y GeoNames utilizando la propiedad *owl:sameAs*.

Como resultado, se obtiene una archivo RDF con los datos descritos con esta ontología y enriquecidos con Wikidata.

5.4 Publicación de datos

Publicar implica: (1) creación del repositorio y carga de datos, (2) inclusión de los metadatos descriptivos asociados al conjunto de datos, y (3) hacerlo accesible. En este paso se realiza la carga y publicación del conjunto de datos según los principios de LOD acuñados por Tim Berners-Lee.

1. Se utilizan URIs para identificar los recursos que se publican. En la sección 5.3 se ha definido el patrón de creación de URIs para este conjunto de datos, de manera que se identifiquen de manera unívoca.

2. Utilizar URIs basados en HTTP para que se puedan localizar y consultar esos recursos, siendo interpretables tanto por humanos como por máquinas. Los URIs definidos establecen la ruta a los recursos utilizando descripciones explícitas de las entidades que componen este conjunto de datos, siendo fácilmente identificables.

3. Se proporcionar información útil y procesable en RDF sobre el recurso al que referencia URI.

4. Incluir enlaces a otros repositorios y promocionar el descubrimiento de información. El conjunto de datos ha sido enriquecido con enlaces a Wikidata y Geonames mediante la propiedad `owl:sameAs`.

Creación del repositorio y carga de los datos

Una vez generado el RDF en el paso anterior, se debe seleccionar dónde almacenarlo. Existen diferentes técnicas para publicar conjuntos datos (ver sección 3.5) y para esta aplicación práctica se ha utilizado RDF4J. Este servidor permite el acceso a los datos, validación de restricciones de integridad, indexación por lotes y razonamiento.

El listado 5.5 representa un extracto del RDF generado en el paso anterior en formato NTriple[2]. A modo de ejemplo, se puede ver en la figura 5.6 un extracto del RDF pero generado en formato RDF/XML, más complejo de interpretar por humanos. Estos archivos se pueden cargar en una instancia del servidor RDF4J en el que previamente se ha creado el repositorio. La tabla 5.4 resumen los datos del repositorio una vez cargado el RDF.

Posteriormente se realizaron las siguientes validaciones sobre el conjunto de datos RDF:

- Los datos RDF fueron validados por el validador W3C RDF[3].

- Se realizó un muestreo de aceptación y revisión manual en varios cientos de registros.

[2]Formato de texto plano para codificar un grafo RDF `Tipe MIME application/n-triples`.

[3]Ver `http://www.w3.org/RDF/Validator`

Tabla 5.4: Sumario del repositorio "agua" cargado en el servidor RDF4J.

ID	agua
Título	Native store with RDF Schema inferencing
Dirección	.../rdf4j-server/repositories/agua
Servidor RDF4J	.../rdf4j-server
Punto SPARQL	.../sparql
Número de tripletas	39386
Vocabularios	8
Número de clases	14
Número de propiedades	41

```
<.../ontologies/water.owl#1.1FugasEnRedTransporte2013>
<.../ontologies/water.owl#enElAnyo>
<.../ontologies/water.owl#2013> .
<.../ontologies/water.owl#1.1FugasEnRedTransporte2013>
<.../ontologies/water.owl#unidad>
<.../ontologies/water.owl#numero> .
<.../ontologies/water.owl#1.1FugasEnRedTransporte2013>
<.../ontologies/water.owl#value> "0" .
<.../ontologies/water.owl#1.1NumAcometidasPorLongitud2011>
<.../ontologies/water.owl#enElAnyo>
<.../ontologies/water.owl#2011> .
<.../ontologies/water.owl#1.1NumAcometidasPorLongitud2011>
<.../ontologies/water.owl#unidad>
<.../ontologies/water.owl#numero/km>
```

Figura 5.5: Ejemplo `N-Triple` generado con extensión `.nt` `Tipe MIME application/n-triples`.

```
<rdf:RDF>
<rdf:Description rdf:about="http://www.khaos.uma.es/
    ontologies/lucentia/ontologies/water.owl#4.2
    EdadMediaDistribucion2014">
<j.0:enElAnyo>2014</j.0:enElAnyo>
<j.0:unidad>
http://www.khaos.uma.es/ontologies/lucentia/ontologies/
    water.owl#nanyos
</j.0:unidad>
<j.0:value>30</j.0:value>
</rdf:Description>
<rdf:Description rdf:about="http://www.khaos.uma.es/
    ontologies/lucentia/ontologies/water.owl#3.6
    RenovacionRedAbastecimiento2014">
<j.0:enElAnyo>2014</j.0:enElAnyo>
<j.0:unidad>
http://www.khaos.uma.es/ontologies/lucentia/ontologies/
    water.owl#tantoPorCiento
</j.0:unidad>
<j.0:value>0</j.0:value>
</rdf:Description>
<rdf:Description rdf:about="http://www.khaos.uma.es/
    ontologies/lucentia/ontologies/water.owl#2.7
    NumAcometidas2011">
<j.0:enElAnyo>2011</j.0:enElAnyo>
<j.0:unidad>
http://www.khaos.uma.es/ontologies/lucentia/ontologies/
    water.owl#numero
</j.0:unidad>
<j.0:value>341</j.0:value>
</rdf:Description>
...
```

Figura 5.6: Ejemplo RDF generado Tipe MIME application/rdf+xml.

- Se implementó un procedimiento que prueba que el número de zonas y subzonas cargadas coincide con los números en la fuente de datos original.

Se evidencia en este punto la falta de validación de aspectos tan relevantes en la calidad de los datos como la precisión, integridad, licencia, relevancia y accesibilidad entre otras.

Inclusión de los metadatos asociados

Una vez creado el repositorio en RDF4J se deben incluir metadatos descriptivos que faciliten su descubrimiento. Se proporciona información sobre la licencia y explotación de los datos mediante `dcterms:license` y `dcterms:rights`.

Accesible

El servidor RDF4J facilita un punto de acceso SPARQL para consultar el conjunto de datos. Este punto de acceso permite la consulta de cualquier sentencia SPARQL tanto por usuarios como por parte de las máquinas.

5.5 Explotación de los datos

Para la explotación del conjunto de datos se habilita un punto de acceso SPARQL. Este facilita el acceso y reutilización del conjunto de datos a usuarios expertos y aplicaciones que se conectan automáticamente para consultar y recuperar la información contenida en el repositorio. Las tecnologías LOD permiten el consumo de datos de una manera diferente a los sistemas tradicionales, por ejemplo, con el uso de consultas federadas que permiten obtener datos externos de bases de conocimiento como Wikidata y GeoNames. El listado 5.7 muestra

```
SELECT ?year ?value ?unit
WHERE {
water:zona6.3 water:hasLengthSupplyNetwork ?length .
?length water:inYear ?year .
?length water:unit ?unit .
?length water:value ?value}
ORDER BY ?year
}
```

Figura 5.7: Sentencia SPARQL que recupera el tamaño de la red de suministro de agua por zona y año.

un ejemplo de consulta sencilla en SPARQL que recupera la longitud de la red de suministro de agua por zona y año, datos almacenados en el repositorio.

El listado 5.8 muestra un ejemplo de consulta más compleja que recupera los datos del agua suministrada y la población en una zona particular `Zone4.1` (ver los resultados en la figura 5.9). Estos datos provienen de repositorios distintos, el agua suministrada es información almacenada en el repositorio y sin embargo, la población se recoge de Wikidata en la misma sentencia a través de la instrucción `SERVICE`. Combinando estos datos se obtiene información más completa que facilita la identificación de posibles relaciones entre el agua suministrada y la población en una determinada zona. Con ellos, la empresa suministradora de agua dispone de más información de valor y en un contexto, que le facilita la toma de decisiones y el descubrimiento de relaciones entre medidas que no se habían tenido en cuenta hasta el momento.

Posteriormente se hizo un estudio para comprobar si existía una relación entre el agua suministrada y la población de una zona. Para ello se ha estudiado la correlación entre el agua suministrada y la población, aunque no se disponen de muchos datos en este caso

```
SELECT ?waterSupplied ?population ?wikidataLink
WHERE {
water:Zone4.1 water:hasWaterSupplied ?waterSupplied .
water:Zone4.1 owl:sameas ?wikidataLink.
?waterSupplied water:inYear ?year .
BIND(concat("", ?year) as ?yearTr)
?waterSupplied water:value ?value .

FILTER(regex(str(?wikidataLink), "wikidata"))

SERVICE <https://query.wikidata.org/sparql> {
    ?wikidataLink p:P1082 ?populationStatement.
    ?populationStatement ps:P1082 ?population;
                         pq:P585 ?date.
    FILTER(xsd:integer(YEAR(?date)) = xsd:integer(?
        yearTr))
}
}
```

Figura 5.8: Sentencia SPARQL que recupera el agua suministrada y la población en la zona `Zone4.1`. Los datos de la población se obtienen desde Wikidata.

Tabla 5.5: Resultados de la sentencia 5.8.

waterSupplied	population	wikilink
water:4.2AguaSuministrada2010	334418	http://www.wikidata.org/entity/Q11959¿
water:4.2AguaSuministrada2011	334329	http://www.wikidata.org/entity/Q11959¿
water:4.2AguaSuministrada2012	334678	http://www.wikidata.org/entity/Q11959¿

de estudio. Se utilizó la correlación de Pearson [Pearson, 1896] para medir el grado de variación entre ambas variables. El coeficiente de correlación $R(5)$ fue $-0,6262$, lo que significa una correlación negativa moderada. Analizando si la correlación era significativa con el valor $p = 0,1326$ (mayor que los niveles de significancia estadística de $0,05$ y $0,10$) se concluyó que el resultado no era estadísticamente significativo. Además, se calculó la correlación de Spearman [Spearman, 1904] obteniendo un coeficiente de $Rs = -0,5371$ y $p = 0,2152$ concluyendo que la relación entre las dos variables (agua suministrada y población) no se podía considerar estadísticamente significativa [Hauke and Kossowski, 2011].

Basado en el mismo conjunto de datos utilizado en esta experimentación, se realizó un estudio [Maté et al., 2016] centrado en el análisis de relaciones potenciales entre indicadores y el descubrimiento de objetivos que pudieran estar ocultos. En él, los datos obtenidos no respaldan una relación (inicialmente esperada) entre fugas y roturas en la red, y la pérdida de agua. Después de la experimentación, los autores sugirieron la conveniencia de revisar la forma que monitoriza sus objetivos, es decir, cómo la empresa mide las averías o revise la idoneidad de la relación, esto es, las averías no causan pérdidas graves de agua.

5.6 Valoración mediante criterio de expertos

En esta sección se describen los procedimientos aplicados para poder obtener una valoración del modelo propuesto, pero desde un punto de vista teórico, mediante el criterio de un grupo de expertos. Antes de analizar los resultados obtenidos de esta valoración, se presentan algunos de los datos que avalan la selección de este grupo.

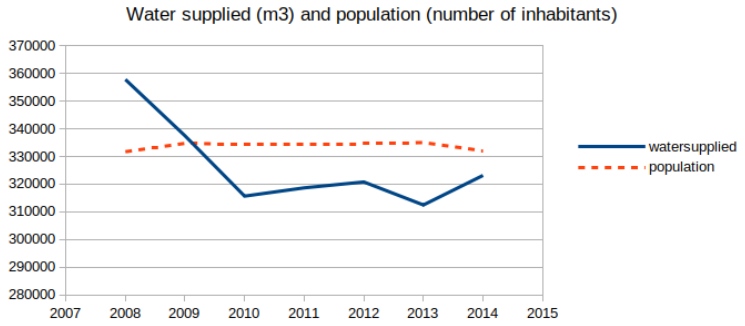


Figura 5.9: Comparativa de suministro de agua y población de la zona 4.1. de 2008 a 2014.

Para la valoración del modelo se seleccionaron 15 expertos basándose en su vinculación y experiencia con la temática tratada en esta tesis y todos aceptaron participar. Primero se calcula el índice de competencia experta K que permite obtener el grado de idoneidad de estos expertos para realizar la valoración del modelo, la tabla C.2 (anexo C) muestra los valores obtenidos por cada experto. Catorce de los expertos han obtenido un índice K mayor o igual a $0,8$ solo uno de los expertos obtuvo un valor por debajo de $0,8$ en concreto $0,75$. Con estos datos, los 15 han sido considerados adecuados para realizar la valoración del modelo, 14 expertos con un coeficiente de competencia alto y 1, con un coeficiente medio. Diez de los expertos son doctores en informática ($66,6\,\%$) y 14 de ellos realizan tareas de docencia e investigación. La muestra recoge expertos de edades comprendidas entre 27 y 63 años y el 13,3 % son mujeres. El proceso completo para el cálculo del índice de competencia experta se detalla en el anexo C.

Para realizar dicha valoración se les suministra información sobre el modelo y las fases que lo conforman en esta primera aplicación práctica (iteración 1), así como los aspectos fundamentales que debe

cumplir. Además, se realizaron entrevistas individuales con el fin de poder aclarar las dudas surgidas sobre el modelo o aspectos más técnicos. Los datos de valoración de los expertos se recogieron a través de una encuesta (ver anexo F).

Para procesar los resultados obtenidos se ha utilizado la escala de Likert [Fabila Echauri et al., 2013], técnica que permite medir el grado de aceptación del modelo por parte de los expertos. Primero se identifica el número de respuestas y frecuencia de cada uno de los niveles de la escala de Likert definida para la encuesta (diseñada para identificar el grado de acuerdo con cada una de las cuestiones planteadas). Después se calcula el índice porcentual IP, que informa sobre el grado de aceptación del grupo de expertos sobre el indicador en cuestión. Se puede consultar el anexo E para obtener más detalles sobre el proceso de cálculo del IP.

Como se observa en la tabla E.1 (anexo E) los cuatro primeros indicadores hacen referencia a los principios fundamentales asociados al modelo. Se ha obtenido una buena aceptación (por encima del 86 %) en la interoperabilidad, reutilización y enriquecimiento (1,2 y 4, respectivamente). Sin embargo, en el tercer indicador (calidad de los datos) se ha obtenido un 40 % que junto con los comentarios suministrados por los expertos, marcan la necesidad de mejorar este aspecto en el modelo dado que no se realiza una evaluación exhaustiva de calidad de los datos y solo se realiza una verificación en el paso de publicación. Con respecto a la valoración de los cuatro pasos que componen el modelo en esta iteración, los tres primeros obtuvieron una buena aceptación (por encima del 90 %). Sin embargo, la fase de explotación obtuvo un grado de aceptación menor (70,67 %), quizás afectado por la falta de comprobación de la confiabilidad y consistencia de los datos publicados y que van a ser explotados. Además, destacan la falta de un visualización de los datos más sencilla, orientada a usuarios menos expertos en la materia (ya que no todos conocen cómo interrogar un repositorio LOD con SPARQL) y dicho sea de paso, problema muy

comentado en la literatura [Bianchini et al., 2018, Rani et al., 2017]..

El procesamiento realizado junto con los comentarios de los expertos evidencia la necesidad de trabajar el aspecto fundamental de la calidad de los datos junto con nuevas formas de visualizar los datos que repercutirá de manera directa en la reutilización y explotación de los datos. El modelo en esta iteración, no ofrece la suficiente confiabilidad en los datos que genera para que puedan ser reutilizados y explotados con total confianza. Los expertos coinciden en que el modelo no contempla de manera adecuada la evaluación de los datos en sí mismos, no ofrece un mecanismo claro para la comprobación de los datos. Esta fase de comprobación daría un aspecto más formal y fiable al modelo presentado. Es importante, en este punto en el que se plantea un refinamiento del modelo, hacer hincapié en la importancia de suministrar datos de calidad contrastada, fiables, para que puedan ser utilizados en procesos como la toma de decisiones y que en definitiva puedan ser útiles para el usuario final, experto o no experto.

Las conclusiones extraídas en esta valoración han dado lugar a una segunda iteración en la que se realiza el refinamiento del modelo. Destacando entre las mejoras llevadas a cabo, la incorporación de un nuevo paso en el modelo para evaluar la calidad de los datos que se generan y se ponen a disposición del público. Además, se contempla una forma de visualizar los datos más sencilla y orientada a usuarios no expertos. La nueva propuesta consta de 6 pasos: (1) especificación de las fuentes datos y preprocesamiento, (2) modelado, (3) generación, (4) publicación, (5) evaluación de la calidad y, (6) explotación. La aplicación de la nueva propuesta del modelo se describe en el capítulo 6 y al igual que se ha hecho en esta iteración, se evalúa mediante el criterio de un grupo de expertos.

6. Aplicación del modelo. Iteración 2

> Dime algo y lo olvidaré,
> enséñame algo y lo recordaré,
> hazme partícipe de algo y lo
> aprenderé.
>
> *Confucio*

En este capítulo se describe la aplicación del modelo ya refinado, después de evaluar la primera iteración y concluir en ella la necesidad de incorporar un nuevo paso para la evaluación de la calidad de los datos y mejorar la parte de explotación del conjunto de datos. La nueva propuesta está formada por los siguientes pasos: (1) especificación de las fuentes datos y preprocesamiento, (2) modelado, (3) generación, (4) publicación, (5) evaluación de la calidad, y (6) explotación.

En esta segunda iteración, se ha incluido un nuevo paso para la evaluación de la calidad basado en la propuesta de [Färber et al., 2018], en la que se propone una lista de criterios de calidad de datos para evaluar los KG en el contexto LOD. Este enfoque utiliza los conceptos de criterios, dimensiones y categorías originalmente propuestos por investigaciones previas sobre calidad de datos [Wang and Strong, 1996].

Para evaluar esta nueva propuesta se aplica el modelo refinado a un escenario real en el contexto de los datos abiertos de la ciudad de Barcelona, en concreto recolectados de su plataforma *Open Data BCN*.

Esta selección está formada por un conjunto de datos heterogéneos en su forma y formato[1]. La figura 6.1 muestra el modelo resultante aplicado a este caso real.

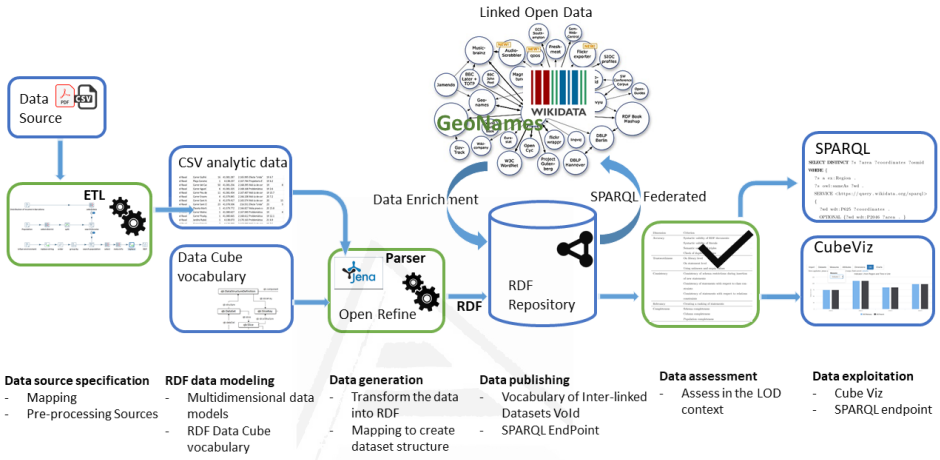


Figura 6.1: Aplicación del modelo refinado (iteración 2) al caso real *Open Data BCN* para la publicación de LOD.

6.1 Introducción

Según el informe sobre el estado de las ciudades europeas[2], y las prioridades establecidas para impulsar el desarrollo regional y urbano de

[1] Datos recolectados en enero de 2017.

[2] `https://ec.europa.eu/regional_policy/en/policy/themes/urban-development/cities-report`

la *Unión Europea* (UE)[3] las ciudades de la UE están a la cabeza con respecto a la acción climática, impulsando la innovación y reduciendo el impacto en el planeta.

Se trabaja mucho para sensibilizar al público en general con el objetivo de mejorar la calidad de vida de los ciudadanos. Al explotar y reutilizar los datos proporcionados por las plataformas *Open Data* es posible realizar previsiones, adelantándose a posibles problemas y minimizando su impacto, incluso llegar a evitarlos. Los datos se pueden enriquecer mediante diferentes repositorios y podrían visualizarse en cuadros de mandos que permitan a los responsables de la toma de decisiones analizar las partes críticas de su organización. Sin embargo, una plataforma de datos abiertos puede tener esta información como contenido textual pero no siempre de forma estructurada, lo que dificulta el proceso de búsqueda y su procesamiento.

El modelo propuesto en este trabajo de tesis ha sido evaluado utilizando los datos de la plataforma *Open Data BCN* en el contexto de *Open Government Data*. Este escenario se centra en el estado de los puntos críticos de limpieza en esta ciudad[4]. Un punto de limpieza crítico es una localización en la que se pueden identificar varios problemas, como contenedores de basura llenos, en mal estado, muebles pesados, objetos viejos, etc. Los datos se obtienen de una campaña de comunicación para mejorar la limpieza en la ciudad de Barcelona, que se llevó a cabo en febrero de 2017.

Los detalles de cada paso del proceso de publicación se describen en las siguientes secciones.

[3] `https://ec.europa.eu/regional_policy/en/policy/how/priorities`

[4] `http://opendata-ajuntament.barcelona.cat/data/en/dataset/punts-critics-neteja-barcelona`

6.2 Especificación de las fuentes de datos

En este paso se realiza la especificación e integración de las fuentes de datos de acuerdo con las pautas marcadas en la definición del modelo (sección 4.2): (1) selección de las fuentes, conexión y recolección; (2) análisis y preproceso (preparación y normalización); y (3) generación de una vista integrada. Como resultado se obtiene un fichero CSV. Es un formato de texto muy utilizado para el almacenamiento de datos tabulares y cuyos campos se separan utilizando, en este caso, una coma. Este es un paso semiautomático que requiere de un análisis previo y de la identificación de puntos comunes que permitan la integración de las diferentes fuentes de datos.

Selección de las fuentes, conexión y recolección

En este caso real se utilizan fuentes de datos de administraciones públicas, para la selección adecuada de las fuentes se han seguido estos dos pasos:

- Se reutilizan datos abiertos publicados por la plataforma *Open Data BCN*[5]. La tabla 6.1 lista los conjuntos de datos seleccionados para esta experimentación, junto con el formato en el que estaban disponibles originalmente en la fecha en la que se realizó esta prueba.

- Se identifican los conjuntos de datos que comparten puntos comunes (distrito, ubicación geográfica, etc.) y por tanto, permiten un análisis más detallado.

[5] `http://opendata-ajuntament.barcelona.cat`

Tabla 6.1: Conjuntos de datos gubernamentales utilizados en el proceso de transformación.

Datos	procedencia	formato	
Entorno urbano	Open Data BCN	hoja de cálculo	CSV
Límites administrativos	Open Data BCN	hoja de cálculo	CSV
Distribución de ingresos en Barcelona	Open Data BCN	texto	PDF
Población	Open Data BCN	texto	PDF

En primer lugar, se realizó un análisis de los datos disponibles en la plataforma *Open Data BCN*, inicialmente se seleccionaron archivos sobre entorno urbano y estado de las áreas críticas de limpieza en la ciudad de Barcelona así como información sobre límites administrativos. Estos datos están disponible para su descarga en formato CSV. Posteriormente, se recolectaron los datos sobre la distribución de los ingresos y la población en la ciudad de Barcelona. En el momento de la recolección estaban disponibles en formato PDF[6].

Análisis y preproceso

Una vez recolectados los datos se analizan. Del fichero de entorno urbano se extraen los datos sobre los puntos críticos relativos a la limpieza de la ciudad de Barcelona. También se incluye información sobre la ubicación geográfica, el vecindario, las visitas realizadas a un determinado punto crítico, en algunos casos, se indica la razón por la

[6]Posteriormente a esta experimentación la plataforma Open Data Barcelona renovó su portal ofreciendo sus datos en formatos procesables automáticamente como por ejemplo CSV. Actualmente en su catálogo no hay ningún *dataset* con recursos en formato PDF (visitada en marzo de 2020).

```
Martina Castells ;1;41.379772;2.166827;Mala praxis comercial;20;15.8;18;7.4;18;10.8;11;4.3;18;4.2
Malnom ;1;41.380627;2.167949;Problemàtica social;19;4;18;4.9;18;3.3;16;4.6;19;3.7
Picalquers ;1;41.380665;2.168612;Problemàtica social;19;12.1;18;8;18;12.7;16;13.1;19;12.3
Rubió i Luch;1;41.38073;2.170163;Problemàtica social;21;6.9;19;4.3;18;4.7;16;3.2;19;2.9
Egipcíaques;3;41.38089;2.168989;Mal ús de contenidors i/o papereres;19;7.6;19;9.3;18;10.4;16;7.3;20;6.9
Gardunya ;1;41.381252;2.17052;Força afluència de gent;19;7.1;19;5.8;17;5.3;16;4.1;19;5.4
Sant Agustí;0;41.38073;2.171528;Mala praxis comercial;19;9.2;19;14.6;18;6.2;16;6.7;19;11.9
Robador;53;41.378761;2.171453;Efecte crida;23;3.4;19;3.6;18;5.8;16;6.4;20;8.3
Salvador Seguí;1;41.37878;2.170497;Mal ús de contenidors i/o papereres;23;5.3;19;5.5;18;3.6;16;4.1;20;7.6
Aurora ;11;41.378757;2.168208;Efecte crida;23;7.3;18;11.1;18;15.1;16;18.5;19;16.1
Aurora ;25;41.377934;2.167251;Efecte crida;23;10.8;18;7.5;18;16.1;16;15.7;19;18.9
Sant Rafael ;42;41.378326;2.1683;Mal ús de contenidors i/o papereres;23;13.3;17;14.8;18;17.1;16;17.1;19;17.8
```

Figura 6.2: Extracto de los datos que contiene el fichero con información sobre los puntos críticos de limpieza en la ciudad de Barcelona. Datos extraídos de Open Data BCN.

cual se considera que esa localización es un punto crítico. La figura 6.2 muestra un extracto de los datos obtenidos en su formato original. Para validar los datos sobre los barrios y zonas de la ciudad, se ha utilizado otra fuente de datos de la que se ha obtenido información sobre los límites administrativos oficiales de la ciudad. Con estos datos se ha podido cotejar que la información sobre los barrios y las zonas de Barcelona son correctos.

Para analizar si la población y la renta per cápita podrían estar relacionados con los puntos críticos, se ha combinado esta información con el archivo comentado anteriormente sobre el entorno urbano y los límites administrativos. Al contrario que ocurrió con los datos extraídos anteriormente, la distribución de los ingresos y los datos relativos a la población en la ciudad de Barcelona estaban disponibles en formato PDF. Por ello fue necesario extraer previamente el texto de los archivos PDF. En este caso, se situaría en la categoría "una estrella" de la clasificación propuesta por Tim Berners-Lee al publicar los datos en la Web en cualquier formato y bajo una licencia abierta, independientemente de lo difícil que sea procesarlos, como por por ejemplo documentos en formato PDF no legible o imágenes escaneadas.

Una vez preparados los archivos fuente se realiza la integración de

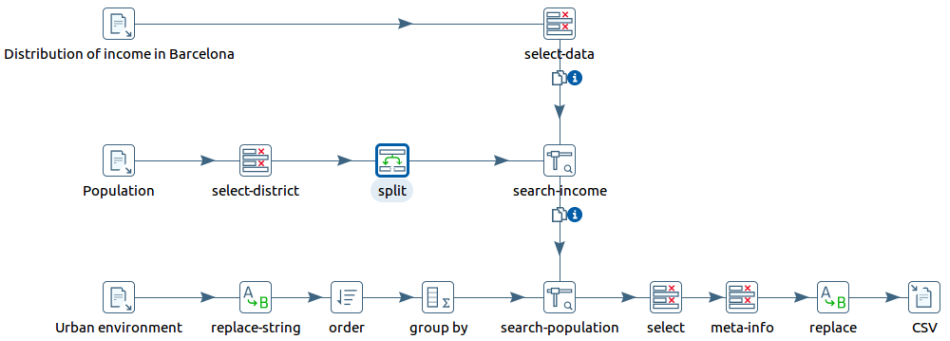


Figura 6.3: Proceso ETL diseñado con Pentaho Data Integration (Kettle).

los datos mediante procesos ETL diseñados mediante la herramienta Data Integration de Pentaho (Kettle). La figura 6.3 muestra una representación gráfica del proceso de transformación realizado. Este proceso tiene tres puntos de entrada que corresponden a tres fuentes de datos heterogéneas en términos de formato y contenido.

El proceso ETL incluye tareas adicionales de limpieza y normalización de datos necesarias debido a que provienen de diferentes archivos y algunos datos no son consistentes entre sí. Algunos de los problemas encontrados en este proceso son: diferentes campos que describen el mismo contenido o un mismo campo contiene información de varios atributos, en este caso sería necesario procesar el texto para extraer los datos por separado (por ejemplo, la siguiente cadena de texto *3.la Barceloneta* contiene el código y el nombre del vecindario).

Generación de una vista integrada

Un vez seleccionadas las fuentes y procesados los datos se genera un archivo CSV utilizado en los siguientes pasos de este proceso de publicación de LOD, en concreto para realizar el modelado y generación del RDF.

6.3 Modelado de datos RDF

Una vez preparados los datos, el siguiente paso es la transformación a un modelo multidimensional que incluye como componentes: dimensiones, medidas y atributos. Para ello es necesario realizar: (1) selección de la ontología adecuada para el modelado del dominio, (2) en caso de no tener una adecuada se creará, priorizando la reutilización de recursos y vocabularios existentes, y (3) diseño del URI.

Selección de la ontología

Siempre que se manejan datos subyace un modelo de datos que lo describe. En el modelo conceptual (abstracto) se definen los datos y sus relaciones, además de posibles restricciones de diferente tipo. Posteriormente se transforma en un modelo ontológico. En este paso se utiliza el archivo CSV obtenido previamente y se realiza una representación conceptual de la estructura del conjunto de datos como un esquema *snowflake*. En él se refleja la organización jerárquica de las dimensiones (normalizadas) y permite una mejor comprensión de los niveles de clasificación definidos en la dimensión.

La figura 6.4 muestra la representación conceptual con el esquema *snowflake*. En ella se puede identificar la tabla de hechos `fact_spots` que almacena los datos agregados obtenidos del paso anterior. En ella se almacenan los puntos críticos de limpieza, número de visitas, ingresos per cápita, población y estado. Alrededor de la tabla de hechos

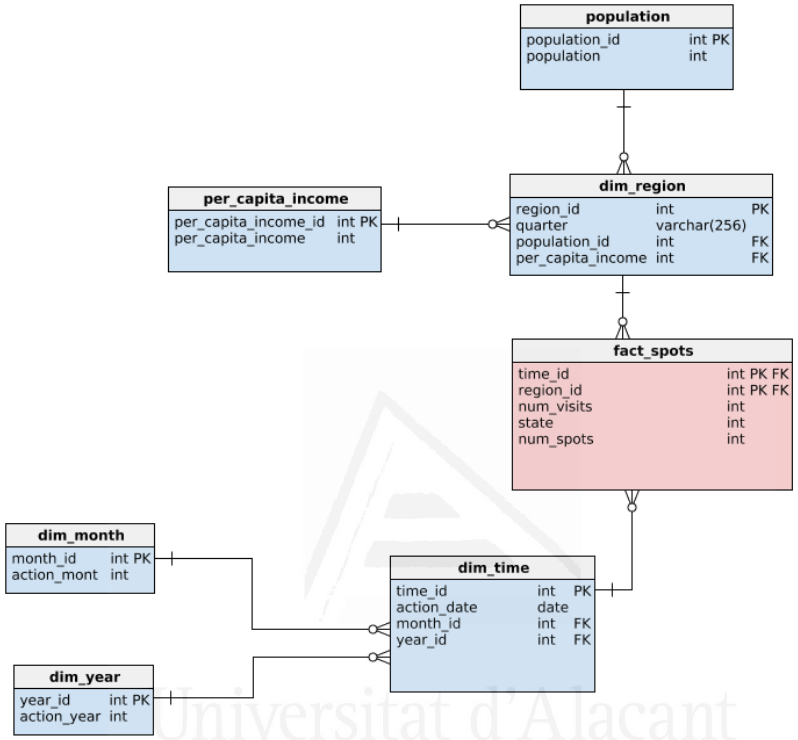


Figura 6.4: Representación conceptual del esquema *snowflake* para los puntos críticos de limpieza en la ciudad de Barcelona.

se sitúan las distintas dimensiones, en concreto en la figura se puede observar las dimensiones relativas al tiempo `dim_time` y a las regiones `dim_region`.

Para publicar los datos multidimensionales se ha seleccionado el vocabulario RDF Data Cube recomendado por el W3C. En él se utiliza la clase *qb:dataSet* para identificar una colección de observa-

ciones definidas como *qb:Observation*. Las dimensiones, los atributos y las medidas se representan como propiedades RDF, definidas por la clase abstracta *qb:ComponentProperty*, que contiene las subclases *qb:DimensionProperty*, *qb:AttributeProperty* y *qb:MeasureProperty*. Como se ha comentado anteriormente, los componentes de una dimensión identifican las observaciones, por ejemplo, el momento en que se produce la observación o su localización. Los componentes de medida representan el hecho que se está observando, mientras que los componentes de atributo permiten la calificación e interpretación de los valores observados al agregar metadatos sobre unidades de medida o el estado de la observación.

La publicación de datos multidimensionales por medio del vocabulario RDF Data Cube aporta varios beneficios importantes. En primer lugar, el conjunto de datos se publica en un formato computacionalmente tratable y no propietario en lugar de proporcionar archivos estáticos como CSV y PDF. Además, las observaciones individuales son direccionables, lo que permite el uso de datos de terceros mediante la creación de referencias. Y por último, hay que tener en cuenta que existen otras herramientas basadas en RDF Data Cube, lo que facilita su reutilización.

Diseño del URI

Siguiendo las pautas de diseño para la publicación de LOD [Tim Berners-Lee, 2006] y el vocabulario RDF Data Cube, el enfoque propuesto se caracteriza por la siguiente estructura:

- El conjunto de datos se identifica por `base_URI/dataset`. Un recurso que representa todo el conjunto de datos se crea y se representa como `qb:DataSet` y después se enlaza con la correspondiente definición de la estructura de datos a través de la propiedad `qb:structure`.

- La definición de la estructura de datos del *dataset* se identifica como `base_URI/dsd` e incluye dimensiones, atributos y medidas. Se representa como `qb:DataStructureDefinition`.

- El vocabulario RDF Data Cube representa las dimensiones, atributos y medidas como propiedades RDF. Cada uno, es una instancia de la clase abstracta `qb:ComponentProperty` que contiene las subclases `qb:DimensionProperty`, `qb:AttributeProperty` y `qb:MeasureProperty`. Por ejemplo, el tiempo y la región se referencian como `qb:DimensionProperty`, mientras que la población, el número de visitas y los puntos críticos de limpieza se referencian como `qb:MeasureProperty`.

- Cada barrio representa un recurso identificado mediante el URI `base_URI/quarterquarter_identifier`. Por ejemplo, el barrio La Sagrada Familia cuyo identificador es 6 (en los datos recolectados), se identifica por el URI `base_URI/quarter6`.

- Para identificar los años y meses utilizados en múltiples conjuntos de datos se utiliza el URI `base_URI/Yyear` para los años y `base_URI/YyearMmonth` para identificar un mes de un año en concreto. Por ejemplo, el año 2017 se define como `Y2017` mientras que `Y2017M1` corresponde a enero de 2017.

- Finalmente, cada observación se describe como `qb:Observación` identificada por un URI que contiene la fecha seguida de un número autoincremento. Por ejemplo, `base_URI/201702/obs1` y `base_URI/201705/obs2`.

La tabla 6.2 enumera los prefijos del espacio de nombres utilizados en el conjunto de datos.

El modelo de cubo de datos obtenido como resultado de este paso está basado en el vocabulario RDF Data Cube, en el que cada recurso

Tabla 6.2: Prefijos de los espacios de nombres utilizados en el *dataset*.

Prefijo	URI
dc	`http://purl.org/dc/elements/1.1/`
dcterms	`http://purl.org/dc/terms/`
foaf	`http://xmlns.com/foaf/0.1/`
gn	`http://www.geonames.org/ontology#`
xmls	`http://www.w3.org/2001/XMLSchema#`
owl	`http://www.w3.org/2002/07/owl#`
qb	`http://purl.org/linked-data/cube#`
rdf	`http://www.w3.org/1999/02/22-rdf-syntax-ns#`
rdfs	`http://www.w3.org/2000/01/rdf-schema#`
skos	`http://www.w3.org/2004/02/skos/core#`
sdmx-meas	`http://purl.org/linked-data/sdmx/2009/measure#`
sdmx-attr	`http://purl.org/linked-data/sdmx/2009/attribute#`
sdmx-concept	`http://purl.org/linked-data/sdmx/2009/concept#`
void	`http://www.w3.org/TR/void#`
wd	`http://www.wikidata.org/entity/`
wdt	`http://www.wikidata.org/prop/direct/`

se identifica mediante un URI de manera que se puedan referenciar fácilmente así como beneficiarse del valor de LOD.

6.4 Generación de datos

El siguiente paso es la generación de los datos en RDF y este paso incluye la definición de un método para dicha transformación: (1) generación de los datos en RDF según la ontología seleccionada en el paso anterior (sección 6.3), y (2) enriquecimiento semántico.

Generación del RDF

En este paso se genera el conjunto de datos en RDF haciendo uso de la herramienta OpenRefine para la transformación de los datos al vocabulario RDF Data Cube. OpenRefine, es una herramienta eficaz para trabajar con datos heterogéneos, transformándolos en un vocabulario uniforme y enriqueciéndolos con repositorios externos. El *mapping* se utiliza para crear la estructura del conjunto de datos junto con las observaciones y componentes, utilizando el URI apropiado para cada elemento.

Lo primero que se realiza en este proceso es la carga del CSV en OpenRefine. Después se edita el esquema RDF, quien marca las reglas de mapeo y relaciona una columna de datos con el campo correspondiente de la ontología, especificando cómo se generarán los datos RDF a partir de los datos suministrados. Las celdas en cada registro de datos se colocan en nodos dentro del esqueleto. Un ejemplo de mapeo consiste en relacionar la columna `código del barrio` con el concepto de la ontología `ex:geo`.

En primer lugar, el recurso que identifica el conjunto de datos `ex:dataset` se escribe como `qb:DataSet` y adicionalmente se pueden proporcionar detalles como una breve descripción y la licencia. Después, se define un recurso `qb:DataStructureDefinition` que hace referencia a un conjunto de recursos `qb:ComponentSpecification`. Cada `qb:ComponentSpecification` hace referencia a una dimensión mediante la propiedad `qb:dimension` o bien a una medida a través de `qb:measure`.

Las dimensiones se escriben como `qb:DimensionProperty` y las medidas como `qb:MeasureProperty`. Por ejemplo, `ex:geo` sería una dimensión (recoge los barrios) mientras que `ex:perCapitaIncome` sería una medida tal y como se puede observar en la figura 6.5.

Después, se definen los años y trimestres que proporcionan enlaces a Wikidata y GeoNames. Finalmente, las medidas y dimensio-

```
ex:perCapitaIncome a rdf:Property, qb:MeasureProperty ;
    rdfs:label "per capita income"@en ;
    rdfs:subPropertyOf sdmx-meas:obsValue ;
    rdfs:comment "The per capita income of a particular...;
    qb:concept sdmx-concept:obsValue ;
    rdfs:range xmls:decimal .

ex:geo a qb:DimensionProperty ;
    rdfs:label "Region"@en .
```

Figura 6.5: Descripción de la medida `ex:perCapitaIncome` y de la dimensión `ex:geo`.

nes se utilizan para describir las observaciones que se describen como `qb:Observation`. La figura 6.6 muestra un ejemplo de la asignación utilizada para crear el conjunto de datos RDF.

Figura 6.6: Esquema RDF alineado con la herramienta OpenRefine.

Enriquecimiento semántico

Para promover la reutilización de datos y la interoperabilidad, los barrios se han enlazado a Wikidata y GeoNames mediante una propiedad `owl:sameAs`. En el listado 6.7 se pueden consultar algunos de ellos, donde `wd` es el prefijo de nombre de Wikidata y `gn` de GeoNames.

Finalmente, desde OpenRefine se pueden exportar los datos a RDF en diferentes formatos como Turtle o RDF/XML. El listado 6.8 muestra un extracto del fichero en formato Turtle. En él se identifica el recurso `ex:quarter1` descrito como `ex:Region` y representa el barrio *El Raval*. El recurso `ex:Y2017M2` representa la fecha *Febrero de 2017* que además está enlazado con Wikidata mediante `ow:sameAs` (la entidad Q1404773 representa el mes de Febrero de 2017 en Wikidata). Y el último elemento del ejemplo `obs0` se describe como `qb:Observation` que incluye propiedades que utilizan los recursos definidos anteriormente, como la región y la fecha.

6.5 Publicación de datos

El siguiente paso después de crear el grafo RDF es publicarlo. La publicación consiste en: (1) la creación del repositorio y carga de los datos RDF, (2) inclusión de los metadatos descriptivos asociados al conjunto de datos, y (3) hacerlo accesible para que pueda ser reutilizado. Este conjunto de datos está marcado como *Creative Commons Attribution 4.0*[7], licencia utilizada para las fuentes de datos gubernamentales.

La publicación del conjunto de datos sigue los principios de LOD acuñados por Tim Berners-Lee.

1. En la sección 6.3 se definen los patrones para el diseño de los URIs que identifican los recursos. Por ejemplo, para el conjunto

[7] `https://creativecommons.org/licenses/by/4.0/`

```
ex:quarter1 a ex:Region ;
  owl:sameAs wd:Q1758503 ;
  owl:sameAs gn:3123673 ;
  rdfs:label "El Raval"@en .

ex:quarter2 a ex:Region ;
  owl:sameAs wd:Q17154 ;
  owl:sameAs gn:6544577 ;
  rdfs:label "Gothic Quarter"@en .

ex:quarter3 a ex:Region ;
  owl:sameAs wd:Q377070 ;
  owl:sameAs gn:3128757 ;
  rdfs:label "La Barceloneta"@en .

ex:quarter4 a ex:Region ;
  owl:sameAs wd:Q2442135 ;
  owl:sameAs gn:3119123 ;
  rdfs:label "Sant Pere, Santa Caterina i la Ribera"@en .

ex:quarter68 a ex:Region ;
  owl:sameAs wd:Q1404773 ;
  owl:sameAs gn:6545114 ;
  rdfs:label "El Poblenou"@en .

ex:quarter7 a ex:Region ;
  owl:sameAs wd:Q1904302 ;
  owl:sameAs gn:6690786 ;
  rdfs:label "Dreta de l'Eixample"@en .

ex:quarter36 a ex:Region ;
  owl:sameAs wd:Q3297889 ;
  rdfs:label "La Font d'en Fargues"@en .
```

Figura 6.7: Ejemplo de enlazado de distintos barrios de Barcelona con entidades en Wikidata y GeoNames

```
ex:quarter1 a ex:Region;
  rdfs:label "El Raval"@en;
  owl:sameAs gn:3123673, wd:Q1758503 .

ex:Y2017M2 a ex:Time;
  owl:sameAs wd:Q23419257;
  skos:broader ex:Y2017Q1;
  skos:notation "Y2017M2";
  skos:prefLabel "2017/february"@en .

<http://example.com/201702/obs0> a qb:Observation;
  ex:criticalCleaningSpots 5.0E1;
  ex:geo ex:quarter1;
  ex:numVisits 1.284E3;
  ex:perCapitaIncome 7.46E1;
  ex:population 4.7274E4;
  ex:state 7.3E0;
  ex:time ex:Y2017M2;
  qb:dataSet ex:dataset;
  sdmx-attr:unitMeasure ex:unit .
```

Figura 6.8: Ejemplo de datos generados en Turtle. El recurso `ex:quarter1` descrito como `ex:Region` representa el barrio *El Raval*, el recurso `ex:Y2017M2` representa la fecha *Febrero de 2017* y `obs0` se describe como `qb:Observation`

de datos su URI es `base_URI/dataset`, o para los barrios su URI se forma como sigue: `base_URI/quarterquarter_identifier`.

2. Se han diseñado URIs descriptivos en función del tipo de entidad que definen, para que los recursos sean fácilmente localizables y se puedan consultar tanto por humanos como por máquinas. Además, están basados en HTTP.

3. Se facilita información útil y procesable en formato RDF sobre

el recurso cuando se desreferencie el URI. Además, se facilita un punto de acceso SPARQL cuyo resultado puede ser interpretado automáticamente.

4. Se han incluido enlaces a Wikidata y GeoNames de manera que los barrios (regiones) están conectados a los datos sobre ellos en estos dos repositorios, ofreciendo una información mucho más completa.

Creación del repositorio y carga de los datos

Como se ha comentado en la sección 3.5, existen diferentes alternativas para almacenar y publicar conjuntos datos, y en esta experimentación se han aplicado dos formas diferentes de publicar los datos.

La primera y más sencilla es utilizar DataHub[8] como plataforma de publicación (ver imagen 6.9). En este caso se ha puesto a disposición de los usuarios dos archivos: `rdfdatacube.ttl` que contiene el grafo RDF en formato Turtle, este se puede cargar directamente en un servidor de RDF para su explotación; y el archivo `dfdatacube-critical-cleaning-spots-bcn_zip`, versión comprimida del *dataset* y que incluye datos CSV y JSON normalizados con datos originales y datapackage.json.

[8] `https://datahub.io/smartdataua/rdfdatacube-critical-cleaning-spots-bcn`

Figura 6.9: Archivos publicados para el repositorio *Rdfdatacube critical cleaning spots BCN* y disponibles para su descarga en `datahub.io/smartdataua/rdfdatacube-critical-cleaning-spots-bcn`.

Además de tener disponible el conjunto de datos para su descarga desde DataHub, se ha utilizado el RDF4J para el almacenamiento y gestión del RDF. El servidor RDF4J está compuesto por una interfaz web para la administración del servidor (rdf4j-workbench) y el propio servidor (rdf4j-server) y ambos módulos están desplegados en el servidor de aplicaciones JBoss EAP 6.4 (conocido actualmente como WildFly)[9]. Además, proporciona un punto de acceso SPARQL para el conjunto de datos.

Haciendo uso del archivo `rdfdatacube.ttl` obtenido en el paso anterior, se crea el repositorio en el servidor y se cargan los datos. Las características más relevantes del conjunto de datos se pueden consultar en la tabla 6.3.

Inclusión de los metadatos asociados

El conjunto de datos está descrito mediante el vocabulario VoID. Este vocabulario ayuda a los productores de datos a publicar metadatos en un formato humano y legible por máquina. Se ha incluido información sobre el título, descripción, editor, licencia y url entre otras. El fichero completo con la descripción del repositorio está disponible en el anexo B.

Accesible

Se proporciona un punto de acceso SPARQL para la consulta y reutilización del conjunto de datos. Además, se incluye el fichero void.ttl para describir y facilitar su descubrimiento.

Se puede concluir que el conjunto de datos publicado es LOD, porque cumple con las 5 estrellas del esquema de clasificación propues-

[9] JBoss es un servidor de aplicaciones de código libre, desarrollado completamente en Java `https://wildfly.org/`.

Tabla 6.3: Resumen de las características más relevantes del repositorio "rdfdatacube" cargado en el servidor RDF4J.

ID	rdfdatacube
Título	Escenario modelo multidimensional
Dirección	.../rdf4j-server/repositories/rdfdatacube
Servidor RDF4J	.../rdf4j-server
Descripción	.../void.ttl
Punto SPARQL	.../sparql
Número de tripletas	5235
Número de contextos	1
Vocabularios	14
Número de clases	28
Número de propiedades	38
Enlaces a través de owl:sameAs	81
Regiones	69
Observaciones	483
Dimensiones	2
Medidas	7

to por Tim Berners-Lee descrito en la sección 3.2. Además, cumple con los principios de: interoperabilidad, reutilización y enriquecimiento semántico. Los datos han sido modelados con una ontología y se han utilizados estándares de web semántica, porque los recursos se identifican con un URI que facilita su localización y consulta y se han incluido metadatos descriptivos que facilitan el descubrimiento del repositorio. Además, el conjunto de datos ha sido enlazado con los repositorios externos Wikidata y GeoNames proporcionando información más completa y relacionada sobre los barrios, con la información contenida en estos repositorios.

6.6 Evaluación de LOD

A continuación se evalúa la calidad de los datos siguiendo la propuesta de [Färber et al., 2018]. Las dimensiones y criterios listados en la tabla 6.4 se definieron inicialmente para los KG. A continuación se detalla el procedimiento para evaluar cada criterio adaptado a las especificidades de los *datasets* multidimensionales. Los resultados obtenidos para los diferentes criterios se pueden consultar al final de este apartado en la tabla 6.7.

Precisión

Esta dimensión mide el grado en el que los datos son correctos, confiables y sin errores [Wang and Strong, 1996]. Es una medida crítica para la calidad de los datos puesto que permite evaluar tanto la validez sintáctica como semántica. La generación y gestión de datos de calidad debe ser un requisito previo para la explotación y posterior intercambio de datos. Los usuarios esperan datos sin errores, precisos y que puedan ser explotados. Esta dimensión se ha evaluado mediante tres criterios:

- *Validez sintáctica de los documentos RDF*. Este criterio viene definido por:

$$m_{\text{synRDF}} = \begin{cases} 1 & \text{todos los documentos RDF son válidos} \\ 0 & \text{en otro caso} \end{cases} \tag{6.1}$$

  Se han validado los documentos con RDF NTriples/Turtle Validator[10] confirmando que todos eran documentos RDF sintácti-

[10] `http://ttl.summerofcode.be/`

Tabla 6.4: Listado de los criterios de calidad clasificados por dimensión y categoría.

Categoría	Dimensión	Criterio
Intrínseca	Precisión	Validez sintáctica de los documentos Validez sintáctica de los literales Validez semántica de tripletas
	Fiabilidad	Integridad a nivel de conjunto de datos Integridad a nivel de instancia Uso de valores vacíos y desconocidos
	Consistencia	Consistencia de las restricciones de esquema durante la inserción de instancias Consistencia de las instancias con respecto a las restricciones de clase Consistencia de las instancias con respecto a las restricciones de relación
Contextual	Relevancia	Crear un ranking de instancias
	Integridad	Integridad de esquema Integridad de columna Integridad de población
	Actualidad	Frecuencia de actualización Especificación del periodo de validez de las instancias Especificación de la fecha de modificación de las instancias
Calidad de los datos representacional	Facilidad de entendimiento	Descripción de los recursos Etiquetas en varios idiomas Serialización comprensible del RDF URIs autodescriptivos
	Interoperabilidad	Evitar nodos en blanco y la cosificación del RDF Proveer diferentes formatos de serialización Uso de vocabulario externo Interoperabilidad del vocabulario propietario
Accesibilidad	Accesibilidad	Posibilidad de desreferenciar recursos Disponibilidad del conjunto de datos Disponibilidad de un punto de acceso SPARQL público Opción de descarga en RDF Negociación de contenido Enlazado del sitio *HyperText Markup Language* (HTML) con la serialización RDF Suministro de metadatos del repositorio
	Licencia	Suministro de información legible por máquina sobre la licencia
	Enlazado	Enlazado mediante `owl:sameAs` Validez de los URIs externos

camente válidos (figura 6.10), por tanto el valor de este criterio es 1.

- *Validez sintáctica de los literales*. Permite evaluar si los valores literales cargados en el conjunto de datos son sintácticamente correctos. El grafo RDF G está definido por tripletas (s, p, o) y un conjunto de literales L y este criterio se define:

$$m_{\text{synLit}} = \frac{|\{G \wedge L \wedge o \,\text{es válido}\}|}{|\{G \wedge L\}|} \tag{6.2}$$

  Con el mencionado RDF NTriples/Turtle Validator ha sido posible verificar los errores de tipo de datos. Además, las propiedades como el tiempo, la latitud y la longitud asociadas con los puntos críticos de limpieza se han verificado mediante expresiones regulares. El valor obtenido para este criterio es $0,9976$.

- *Validez semántica de tripletas*. Este criterio evalúa si las declaraciones descritas por las tripletas son ciertas. Para comprobar su validez se verifica si está disponible en una fuente confiable, por ejemplo en la plataforma oficial *Open Data BCN* o Wikidata. Aplicando la fórmula 6.3 (donde G conjunto de tripletas del *dataset* y S la fuente confiable de referencia), el valor obtenido indica un alto porcentaje de datos correctos. Por ejemplo, el RDF del listado 6.11 muestra algunas de los barrios vinculados a Wikidata por medio de la propiedad `owl:sameAs`[11].

$$m_{\text{semTriple}} = \frac{|G \wedge S|}{|G|} \tag{6.3}$$

[11] *El Raval* está vinculado a `https://www.wikidata.org/wiki/Q1758503`.

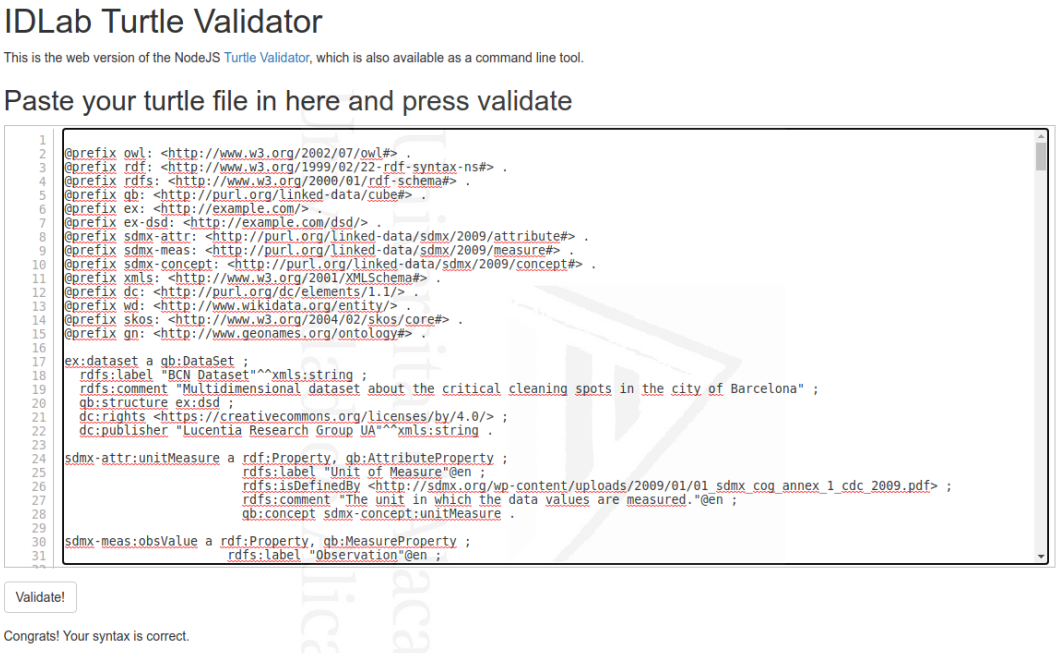


Figura 6.10: Validación sintáctica realizada con la aplicación IDLab Turtle Validator al fichero `rdfdatacube.ttl` obtenido en la sección 6.4.

```
ex:dataset a qb:DataSet ;
rdfs:label "Dataset"^^xmls:string ;
rdfs:comment "Overview description about the dataset" ;
qb:structure ex:dsd ;
dc:publisher "Publisher Office"^^xmls:string .

ex:quarter1 a ex:Region ;
owl:sameAs wd:Q1758503 ;
rdfs:label "El Raval"@en .

ex:quarter2 a ex:Region ;
owl:sameAs wd:Q17154 ;
rdfs:label "Gothic Quarter"@en .
```

Figura 6.11: Extracto de RDF en el que las entidades de tipo `ex:Region` se enlazan a Wikidata mediante la propiedad `owl:sameAs`.

En la dimensión *precisión* se obtienen valores altos en los tres criterios evaluados, por tanto se puede decir que los datos en este caso son correctos, fiables y sin apenas errores.

Fiabilidad

La fiabilidad se define como "el grado en que se acepta que la información es correcta, verdadera, real y creíble" [Zaveri et al., 2016] y se evalúa a tres niveles:

- *Integridad a nivel de conjunto de datos.* Los posibles valores que

puede tener este criterio son:

$$m_{\text{fact}} = \begin{cases} 1 & \text{inserción manual en un sistema cerrado (menos vulnerable)} \\ 0,75 & \text{inserción manual en un sistema mantenido por la comunidad} \\ 0,25 & \text{automatizado, datos extraídos de fuentes de datos estructuradas} \\ 0 & \text{automatizado, datos extraídos de fuentes de datos no estructuradas} \end{cases} \tag{6.4}$$

Como el conjunto de datos se publica mediante una conversión automática a LOD, el valor de este criterio es $0,25$, que corresponde a datos extraídos que provienen de fuentes de datos estructuradas.

- *Integridad a nivel de instancia.* Cumplir con este criterio implica utilizar un vocabulario para describir la procedencia de los datos, además el criterio original distingue entre información de procedencia para tripletas e información de procedencia para recursos. Para describir la procedencia se podría hacer a través de las propiedades `dcterms:provenance` y `dcterms:source` de *Dublin Core* o `prov:wasDerivedFrom` de W3C-PROV [World Wide Web Consortium (W3C), 2013]. Los posibles valores para este criterio son:

$$m_{\text{fact}} = \begin{cases} 1 & \text{información de procedencia a nivel de instancia} \\ 0{,}5 & \text{información de procedencia a nivel de recurso} \\ 0 & \text{en otro caso} \end{cases} \tag{6.5}$$

Como no se ha incluido la descripción del origen de los datos, el valor obtenido en este criterio es 0.

- *Uso de valores vacíos y desconocidos.* No se han utilizado identificadores específicos para capturar aquellos valores desconocidos ni vacíos, por lo que el valor de este criterio es 0 según sus posibles valores en la metodología original:

$$m_{\text{NoVal}} = \begin{cases} 1 & \text{se utilizan valores vacíos y desconocidos} \\ 0{,}5 & \text{se utilizan valores vacíos o desconocidos} \\ 0 & \text{en otro caso} \end{cases} \tag{6.6}$$

El valor obtenido para la dimensión de *fiabilidad* no es muy alto, dado que los datos se recolectan de fuentes de datos abiertos y no se han incluido metadatos para informar sobre su procedencia. Este criterio probablemente debería redefinirse en este contexto, ya que fue creado inicialmente para analizar otro tipo de repositorios. Sería deseable incluir información de procedencia como parte de los metadatos así como identificar valores nulos o desconocidos.

Consistencia

Esta dimensión mide la consistencia para asegurar que dos o más valores dentro del conjunto de datos no entran en conflicto el uno con el otro [Mecella et al., 2002]. La consistencia semántica es la medida en la que las colecciones utilizan los mismo valores y elementos para transmitir los mismos conceptos y significados [Shreeves et al., 2005]. El uso de vocabularios controlados facilita la consistencia del repositorio. En este contexto, OWL permite introducir restricciones para garantizar la consistencia con respecto a clases y relaciones. Esta dimensión se evalúa a través de los siguientes criterios:

- *Consistencia de las restricciones de esquema durante la inserción de instancias.* La comprobación de estas restricciones normalmente se realiza en la interfaz de usuario para evitar inconsistencias durante la inserción de nuevas instancias. Por ejemplo, se podría comprobar que el tipo de entidad es válido cuando se inserta una nueva instancia, esto viene expresado por la propiedad *rdf:type*. Teniendo en cuenta que la interfaz de usuario no realiza restricciones de verificación durante la inserción de nuevas entradas, la puntuación obtenida según los posibles valores (ver 6.7) es 0.

$$m_{\text{checkRestr}} = \begin{cases} 1 & \text{se han comprobado las restricciones de esquema} \\ 0 & \text{en otro caso} \end{cases} \tag{6.7}$$

- *Consistencia de las instancias con respecto a las restricciones de clase.* Mide el grado en el que la instancia es consistente con la clase definida a nivel de esquema. Por ejemplo, la propiedad `owl:disjointWith` se utiliza para validar las restricciones de clase, es decir, a través de esta declaración se puede deducir una inconsistencia. Siendo CC el conjunto de todas las restricciones de clase, definidas como $CC = \{(c_1, c_2)|(c_1, \text{owl:disjointWith}, c_2)\epsilon g\}$ y $c_g(e)$ el conjunto de todas las clases de e en g, definidas como $c_g(e) = \{c|(e, \text{rdf:type}, c)\epsilon g\}$, este criterio se define:

$$m_{\text{conClass}} = \frac{|\{(c_1, c_2)\epsilon CC|\neg\exists e : (c_1\epsilon c_g(e) \wedge c_2\epsilon c_g(e))\}|}{|CC|} \tag{6.8}$$

Estas restricciones se han comprobado a través del punto SPARQL. Por ejemplo, para comprobar que una entidad no es a la vez del

```
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-
    syntax-ns#>
SELECT ?e
WHERE { ?e rdf:type qb:dimension .
        ?e rdf:type qb:measure }
```

Figura 6.12: Sentencia SPARQL que recupera los recursos que sean *Dimension* y *Measure* a la vez.

tipo *Dimension* y *Measure* se ejecuta la consulta de la figura 6.12. En este punto no se han identificado inconsistencias, por tanto, el valor de este criterio es 1.

- *Consistencia de las instancias con respecto a las restricciones de relación.* Este criterio evalúa el grado de consistencia de la instancia con las restricciones de relación y viene definido por:

$$m_{\text{conRelat}} = \frac{1}{n}\sum_{i=1}^{n} m_{conRelat,i}(g) \tag{6.9}$$

Por ejemplo, la relación `rdfs:range` indica el tipo de entidades que pueden ocupar la tercera posición en una tripleta, reduciendo los individuos que una propiedad puede tener como valor. Esta restricción se puede verificar con la consulta SPARQL de la figura 6.13 cuyo resultado para el conjunto de datos actual se visualiza en la figura 6.14.

El valor de la dimensión *consistencia* es relativamente alto, aunque en este caso no se verifiquen las restricciones de esquema durante la inserción de nuevas instancias.

```
PREFIX qb: <http://purl.org/linked-data/cube#>
SELECT distinct ?rangeType
WHERE { ?x qb:measure ?o .
?o a ?rangeType }
```

Figura 6.13: Consulta SPARQL para evaluar el tipo de entidades que pueden aparecer en la tercera posición en una tripleta.

```
RangeType
rdfs:Resource
rdf:Property
qb:MeasureProperty
```

Figura 6.14: Resultado de la consulta SPARQL 6.13 lista el tipo de entidades que pueden aparecer en la tercera posición en una tripleta.

Relevancia

La relevancia es la medida en la que los datos son útiles para una determinada tarea [Zaveri et al., 2016]. Esta métrica se mide a través del criterio *Crear un* ranking *de instancias*:

$$m_{\text{Ranking}} = \begin{cases} 1 & \text{soporta clasificación} \\ 0 & \text{en otro caso} \end{cases} \tag{6.10}$$

El valor de este criterio es 0, dado que el conjunto de datos actual no admite la clasificación de los datos en función de un criterio. En este paso la relevancia se podría aplicar a la hora de mostrar los resultados basándose en la fecha más actual, las zonas con más incidencias o en función de la geolocalización.

Integridad

La integridad es la medida en la que los datos tienen suficiente amplitud, profundidad y alcance para la tarea en cuestión [Wang and Strong, 1996]. Se divide en tres criterios y solo es evaluable mediante un caso de uso concreto. Para valorar esta dimensión es necesario definir un *gold standard*, que es un conjunto de clases y propiedades relevantes para el caso de uso. Las clases y propiedades que lo componen se listan en la tabla 6.5 y están basadas en el vocabulario RDF Data Cube[12] y DBpedia[13].

Tabla 6.5: Gold standard. Clases y propiedades utilizadas para evaluar la integridad.

Clase	Propiedad
Region	name
District	name
Point	name, latitude, longitude
Observation	point, per capita income, population, time, number of visits, state, unit of measure

- *Integridad de esquema.* Se calcula como la proporción del número de clases y atributos del *gold standard* que existe en el conjunto de datos g (no_{clatg}) con respecto a las del *gold standard* (no_{clat}).

$$m_{\text{cSchema}} = \frac{no_{clatg}}{no_{clat}} \tag{6.11}$$

  En este criterio se obtiene una puntuación alta $(0,8)$ porque el vocabulario principal del conjunto de datos evaluado está basado en RDF Data Cube. Sin embargo, en el conjunto de datos no se

[12] `https://www.w3.org/TR/vocab-data-cube/`
[13] `http://dbpedia.org/ontology/`

recopila la latitud y longitud de un punto crítico, pero se podría obtener a través de consultas federadas a Wikidata. Además, esta información se puede encontrar en *Open Data BCN*.

- *Integridad de columna.* Este criterio se define como la tasa de instancias que tienen una propiedad específica definida (no_{kr}), promediada con todas las propiedades en el *gold standard* (no_k), donde H es el conjunto de todas las combinaciones de clases y relaciones.

$$m_{\text{cCol}} = \frac{1}{|H|} \sum_{k,r \epsilon H} \frac{no_{kr}}{no_k} \tag{6.12}$$

  Se ha obtenido un valor de $0,8$ que, como en el criterio anterior, no es el valor más alto porque el conjunto de datos no almacena información sobre los puntos geográficos.

- *Integridad de población.* Mide la cobertura, es decir, hasta qué punto el conjunto de datos abarca la población. E_s representa el conjunto de entidades del *gold standard* y E_g el conjunto de entidades en g. La integridad de población se calcula según:

$$m_{\text{cPop}} = \frac{|E_s \wedge E_g|}{|E_s|} \tag{6.13}$$

  Para calcular esta métrica, se seleccionan cuatro clases y para cada clase dos entidades conocidas *short tail* y dos entidades más desconocidas *long tail*. Para seleccionar las entidades más populares por clase, se utilizan declaraciones cuantitativas. Por ejemplo, para las entidades más conocidas se seleccionan los dos distritos más poblados de Europa. Y para las más desconocidas, dos distritos en el contexto de la ciudad de Barcelona y en un año

específico (2018). La puntuación obtenida es baja $(0,625)$ porque el conjunto de datos carece de entidades conocidas del *gold standard*, aunque todas ellas están representadas en Wikidata.

En esta dimensión el valor obtenido es relativamente alto en el uso de los elementos definidos en el esquema (un resultado natural, en cierta medida, ya que el esquema se ha ajustado a su propósito) y un valor más bajo en la población de datos porque el *dataset* proporciona datos seleccionados basados en la plataforma *Open Data BCN* y no tienen una cobertura universal.

Actualidad

La dimensión *actualidad* es la medida que indica si el dato está suficientemente actualizado para una tarea en cuestión [Pipino et al., 2002]. En esta dimensión se tiene en cuenta si el recurso incluye metadatos sobre cuándo fue creado, almacenado, accedido o citado. Los usuarios esperan que los datos estén actualizados y su frecuencia de actualización es un indicador relevante de calidad [Gonçalves et al., 2007]. Esta dimensión implica la frecuencia e información sobre las actualizaciones llevadas a cabo en el conjunto de datos y se mide a través de los siguientes criterios:

- *Frecuencia de actualización.* Este criterio mide la frecuencia de actualización del conjunto de datos y puede tomar los siguientes

valores:

$$m_{\text{Freq}} = \begin{cases} 1 & \text{actualizaciones continuas} \\ 0{,}5 & \text{actualizaciones discretas y periódicas} \\ 0{,}25 & \text{actualizaciones discretas no periódicas} \\ 0 & \text{no se actualiza} \end{cases} \tag{6.14}$$

Para consultar la frecuencia con la que se actualiza el repositorio se revisan las propiedades como `dcterms:created` y los archivos VoID. Para este conjunto de datos la puntuación obtenida es $0,25$ que corresponde a actualizaciones discretas no periódicas.

- *Especificación del período de validez de las instancias.* Este criterio mide si el repositorio soporta la especificación de rangos o periodos de validez de los datos. En este caso no se usan propiedades como *end time* (`P582`) de Wikidata, para especificar durante cuánto tiempo es válido el conjunto de datos, por lo que el valor de este criterio es 0.

$$m_{\text{Validity}} = \begin{cases} 1 & \text{admite la especificación del período de validez} \\ 0 & \text{en otro caso} \end{cases} \tag{6.15}$$

- *Especificación de la fecha de modificación de las instancias.* Este criterio mide si se utilizan o no las fechas para consultar la última verificación de una instancia representadas por las propiedades `dcterms:modified` o `schema:dateModified`. No se han

detectado estas propiedades en el conjunto de datos, por tanto el valor de este criterio es 0.

$$m_{\text{Change}} = \begin{cases} 1 & \text{especificación de la fecha de modificación de las instancias} \\ 0 & \text{en otro caso} \end{cases} \tag{6.16}$$

Los resultados obtenidos en esta dimensión son bajos puesto que no se han contemplado metadatos para especificar el periodo de validez y modificación de las instancias.

Facilidad de entendimiento

La facilidad de entendimiento indica el grado en que los datos se entienden, son legibles y claros. En este contexto, la dimensión se centra en los usuarios y aborda problemas como el uso de descripciones textuales y URIs descriptivos. Esta dimensión se mide mediante cuatro criterios:

- *Descripción de los recursos.* Los repositorios basados en los principios de Web semántica suelen usar propiedades para describir sus recursos, como por ejemplo, `rdfs:label` y `rdfs:comment`. La fórmula 6.17 define este criterio como el ratio entre las relaciones que contiene la etiqueta o descripción (no_{desc}) y todos los URIs (no_{cr}):

$$m_{\text{Descr}}(g) = \frac{no_{desc}}{no_{cr}} \tag{6.17}$$

  Para este conjunto de datos, la tasa de entidades descritas con la propiedad `rdfs:label` es baja.

- *Etiquetas en varios idiomas.* Este criterio mide si se proporcionan etiquetas en otros idiomas ya que el valor de una propiedad se puede codificar en varios idiomas agregando atributos como `@es`, `@en`, etc.

$$m_{\text{Lang}} = \begin{cases} 1 & \text{proporcionan etiquetas al menos en un idioma adicional} \\ 0 & \text{en otro caso} \end{cases} \tag{6.18}$$

  El conjunto de datos en estudio declara el idioma con estas propiedades `rdfs:label` y `rdfs:comment`, en las que solo se encontraron referencias al inglés, por tanto el valor del criterio es 0.

- *Serialización comprensible del RDF.* Este criterio mide el uso de codificaciones alternativas más comprensibles para los humanos que el RDF/XML (de lectura compleja), como N-Triples, N3 y Turtle [W3C, 2011].

$$m_{\text{uSer}} = \begin{cases} 1 & \text{ofrece serialización del RDF distinta de RDF/XML} \\ 0 & \text{en otro caso} \end{cases} \tag{6.19}$$

  El conjunto de datos proporciona la serialización con Turtle aunque se pueden obtener otros formatos mediante el punto SPARQL[14], por ello el valor del criterio es 1.

- *URI autodescriptivos.* Este criterio evalúa si se utilizan URIs autodescriptivos (descripción legible de la entidad) ayudando así

[14] `http://docs.rdf4j.org/rest-api/#_content_types`

al usuario a comprender mejor el recurso, en lugar de identificadores.

$$m_{\text{uURI}} = \begin{cases} 1 & \text{utiliza URIs autodescriptivos} \\ 0{,}5 & \text{no siempre usa URIs autodescriptivos} \\ 0 & \text{en otro caso} \end{cases} \tag{6.20}$$

El conjunto de datos contiene una descripción y el identificador del recurso por lo que el valor de este criterio es 1.

Los valores que miden la facilidad de comprensión son diversos y dependen del criterio que se evalúe, obteniendo valores bajos en los dos primeros criterios en cuanto a la descripción de los recursos se refiere. Por otra parte, se obtiene un valor alto en lo relativo a los formatos y las URIs autodescriptivas.

Interoperabilidad

La interoperabilidad permite el intercambio de información, datos y conocimiento entre sistemas de forma automática. La interoperabilidad es un aspecto clave para facilitar el intercambio y la reutilización de LOD, por tanto, es importante proporcionar metadatos legibles por máquina [World Wide Web Consortium (W3C), 2005]. Esta dimensión implica dos criterios:

- *Evitar nodos en blanco y la cosificación del RDF.* Este criterio

puede tomar los siguientes valores:

$$m_{\text{Reif}} = \begin{cases} 1 & \text{evita nodos vacíos y la cosificación del RDF} \\ 0{,}5 & \text{tiene nodos vacíos o cosificación del RDF} \\ 0 & \text{en otro caso} \end{cases} \tag{6.21}$$

Como no se han detectado nodos vacíos (verificados mediante el operador de SPARQL, `isBlank`) y el conjunto de datos no utiliza el vocabulario de cosificación del RDF[15], el valor de este criterio es 1.

- *Facilitar diferentes formatos de serialización.* Este criterio mide si se da soporte para otros formatos adicionales a RDF/XML. Los posibles valores de este criterio son:

$$m_{\text{iSerial}} = \begin{cases} 1 & \text{soporta otros formatos además de RDF/XML} \\ 0{,}5 & \text{solo RDF/XML} \\ 0 & \text{en otro caso} \end{cases} \tag{6.22}$$

Se puede configurar la petición para que el servidor RDF proporcione resultados en RDF/XML, JSON-LD y Turtle. Por tanto, el valor de este criterio es 1 con respecto a las especificaciones de [Färber et al., 2018].

- *Uso de vocabulario externo.* Es la proporción de tripletas que utilizan vocabularios externos en el predicado (no_{extVoc}) y el número

[15] `https://www.w3.org/TR/rdf-schema/#ch_reificationvocab`

total de tripletas en el conjunto de datos (no).

$$m_{\text{extVoc}} = \frac{no_{extVoc}}{no} \tag{6.23}$$

El valor obtenido por este criterio es $0,57$ y se ha calculado en función de 24 propiedades de 8 vocabularios externos utilizados en el conjunto de datos. Las propiedades se listan en la tabla 6.6.

Tabla 6.6: Propiedades de vocabularios externos utilizadas en el conjunto de datos.

sdmx-attr:unitMeasure	qb:structure	rdfs:isDefinedBy
qb:attribute	skos:prefLabel	dc:rights
sdmx-meas:obsValue	skos:notation	rdfs:seeAlso
rdfs:domain	rdfs:subClassOf	qb:dimension
qb:dataSet	qb:component	rdfs:range
qb:measure	owl:sameAs	rdfs:subPropertyOf
qb:concept	rdfs:label	rdfs:comment
rdf:type	dc:publisher	skos:broader

- *Interoperabilidad del vocabulario propietario*. Este criterio determina la proporción de clases y propiedades con al menos un enlace de equivalencia a clases y propiedades de vocabularios externos a través de propiedades `owl:sameAs`, `owl:equivalentClass`, `rdfs:subPropertyOf` o `rdfs:subClassOf`.

  Si $P_{eq} = \{$owl:sameAs, owl:equivalenClass, rdfs:subPropertyOf, rdfs:subClassOf$\}$ y U_g^{ext} lo conforman todos los URIs en U_g que son externos al conjunto de datos g, el criterio se mide como sigue:

$$m_{\text{propVoc}} = \{(x, p, o)\epsilon g \wedge (p\epsilon P_{eq} \wedge o\epsilon U_g^{ext})\} \tag{6.24}$$

En este caso las clases y propiedades provienen de vocabularios externos basados principalmente en qb[16], *Dublin Core*, RDF y SKOS, por tanto el valor del criterio es 1.

De la evaluación de esta dimensión se puede concluir que la interoperabilidad del repositorio es alta, proporciona varios formatos de salida y se utilizan vocabularios externos. En este sentido, utilizar vocabularios comunes mejora la interoperabilidad y facilita la integración de los datos.

Accesibilidad

La accesibilidad es la medida en la que los datos están disponibles o son fácilmente recuperables [Wang and Strong, 1996], requiere de un punto SPARQL y que se puedan descargarse los RDF. Además, desde un punto SPARQL pueden ejecutar consultas federadas que permite conectar diferentes conjuntos de datos desde la misma *query*, mejorando y aumentando la visibilidad del LOD. Para medir esta dimensión se tienen en cuenta varios criterios:

- *Posibilidad de desreferenciar recursos.* La desreferenciación de recursos se basa en URIs que se pueden resolver mediante solicitudes HTTP y devuelven información útil y válida. Se considera que ha tenido éxito cuando devuelve el código de estado `http 200` y un documento RDF.

$$m_{\text{Deref}} = \frac{|\text{desreferenciable}(U_g)|}{|U_g|} \tag{6.25}$$

Dado un conjunto de URIs (se seleccionaron 100 de manera aleatoria) y utilizando el campo `application/rdf+xml` en su encabezado HTTP, se comprobó que todos devolvieron un documento RDF correcto, por ello el valor del criterio es 1.

[16] `http://purl.org/linked-data/cube#`

- *Disponibilidad del conjunto de datos.* Este criterio evalúa la disponibilidad del conjunto de datos relativo a tiempos de actividad. Y para ello se calcula el ratio entre el número de peticiones correctas (no_{aval}) y el total de las realizadas (no):

$$m_{\text{Avai}} = \frac{no_{avai}}{no} \tag{6.26}$$

  En este caso se monitorizó el punto SPARQL durante un período de 15 días con intervalos de verificación cada 5 minutos y no se observaron interrupciones en el servicio, por tanto, el valor de este criterio es 1.

- *Disponibilidad de un punto de acceso SPARQL público.* Este criterio hace referencia a la disponibilidad de un punto SPARQL público, de manera que si existe el punto SPARQL y es público el valor del criterio es 1.

$$m_{\text{SPARQL}} = \begin{cases} 1 & \text{punto SPARQL público} \\ 0 & \text{en otro caso} \end{cases} \tag{6.27}$$

  El conjunto de datos creado se almacena en un servidor RDF4J[17] en el que se ha habilitado un punto SPARQL público[18].

- *Opción de descarga en RDF.* Además del punto SPARQL, se puede proporcionar una descarga de los datos en RDF. Este criterio se define:

$$m_{\text{Export}} = \begin{cases} 1 & \text{proporciona la descarga de los} \\ & \text{datos en RDF} \\ 0 & \text{en otro caso} \end{cases} \tag{6.28}$$

[17] `http://rdf4j.org/`

[18] `data.pre.cervantesvirtual.com/rdf4j-server/repositories/rdfdatacube`

El valor del criterio es 1, dado que el conjunto de datos se puede descargar en N-Triples.

- *Negociación de contenido.* Este criterio evalúa la consistencia entre el formato de serialización RDF solicitado y el que devuelve realmente. Los posibles valores son:

$$m_{\text{Negot}} = \begin{cases} 1 & \text{soporta negociación de contenido y retorna el tipo de contenido correcto} \\ 0{,}5 & \text{soporta negociación de contenido pero retorna el tipo de contenido incorrecto} \\ 0 & \text{en otro caso} \end{cases} \tag{6.29}$$

Se ha verificado la coherencia entre el formato de serialización RDF solicitado y el recibido para el caso de RDF/XML, N3, Turtle y N-Triples. Después de validar la respuesta, resultó no ser compatible para Turtle. Por ello, el valor obtenido en este criterio es $0,5$ porque admite la negociación de contenido pero en algún caso devuelve un tipo de contenido incorrecto.

- *Enlace del sitio HTML con la serialización RDF.* El HTML se pueden vincular a la serialización en RDF añadiendo <`link rel=alternate type={content type} href={URL}`> al encabezado. Este criterio toma el valor:

$$m_{\text{HTMLRDF}} = \begin{cases} 1 & \text{patrón de descubrimiento automático usado al menos una vez} \\ 0 & \text{en otro caso} \end{cases} \tag{6.30}$$

No se ha desarrollado un sitio web HTML por el que navegar los contenidos, en consecuencia el valor del criterio es 0.

- *Suministro de metadatos del repositorio.* Este criterio puede tomar estos dos valores en función de si incluye o no metadatos descriptivos:

$$m_{\text{Meta}} = \begin{cases} 1 & \text{metadatos disponibles} \\ 0 & \text{en otro caso} \end{cases} \tag{6.31}$$

  El conjunto de datos se puede describir utilizando VoID [World Wide Web Consortium (W3C), 2011]. En este caso, se ha incluido un archivo VoID con el título, descripción, creadores y vocabularios utilizados (ver anexo B).

En general, los valores para esta dimensión son altos puesto que se proporciona un punto de acceso SPARQL público, el conjunto de datos está disponible para su descarga y funciona sin interrupciones significativas.

Licencia

La licencia se define como la concesión de permiso a un consumidor para que reutilice un conjunto de datos en unas condiciones definidas. Proporcionar una licencia clara y abierta es fundamental para promover la reutilización de datos. La licencia se puede proporcionar por ejemplo, como un texto incluido en el sitio web oficial de la institución o también, en el propio conjunto de datos mediante metadatos legibles por máquina. Esta dimensión se mide con el siguiente criterio:

- *Suministro de información legible por máquina sobre la licencia.* Se puede especificar una licencia por medio de las relaciones

`dcterms:license` y `dcterms:rights` incluidas en el conjunto de datos o en un archivo VoID.

$$m_{\text{macLicense}} = \begin{cases} 1 & \text{dispone de información de licencia legible por máquina} \\ 0 & \text{en otro caso} \end{cases} \tag{6.32}$$

En esta experimentación los datos se distribuyen bajo una licencia Creative Commons[19] y se especifica mediante la propiedad `dcterms:rights`. Además, ofrece el archivo `void.ttl` con los metadatos tanto descriptivos como estructurales sobre el repositorio. El valor obtenido en este criterio es 1 dado que se especifica la licencia.

Enlazado

El enlazado es la medida en la que las entidades que representan un mismo concepto están vinculadas entre sí, ya sea dentro del mismo conjunto o entre dos o más fuentes de datos. Este enlazado es clave para enriquecer un conjunto de datos porque interconectando un conjunto de datos con repositorios externos, se puede crear un nuevo conocimiento. Por ejemplo, si se crea un enlace a GeoNames proporciona un conocimiento bien definido y fiable. Esta dimensión mide el número y la validez de los enlaces externos.

- *Enlazado mediante owl:sameAs.* Este valor se obtiene de la proporción de instancias que tienen al menos un `owl:sameAs` apuntando a un recurso externo. Vinculado con el cuarto principio de *Linked Data* el cual especifica que se deben incluir enlaces a otros repositorios y promocionar el descubrimiento de más información. Y con uno de los principios fundamentales del modelo

[19] `https://creativecommons.org/licenses/by/4.0/`

propuesto en este trabajo de tesis *enriquecimiento semántico* como la interconexión entre los datos que se generan y las bases de conocimiento. Siendo I_g el conjunto de instancias en g se define:

$$m_{\text{Inst}} = \frac{|\{x \epsilon I_g | \exists (x, \text{sameAs}, y) \epsilon g\}|}{|I_g|} \tag{6.33}$$

No se obtiene un valor alto para este criterio, dado que para esta experimentación solo se han enlazado los barrios y años.

- *Validez de los URIs externos*. El enlace a recursos externos puede conllevar que estos dejen de ser válidos con el paso del tiempo. Dada una lista de URIs, este criterio verifica si se produce un *timeout* o un error. Siendo A un conjunto de URIs externos, entonces:

$$m_{\text{URIs}} = \frac{|\{x \epsilon A \wedge x \,\text{es resoluble}\}|}{|A|} \tag{6.34}$$

Para calcular el valor de este criterio se validaron todos los URIs definidos con la relación `owl:sameAs`, al no producirse errores ni *timeout*, el valor del este criterio es 1.

Después de evaluar esta dimensión se evidencia un bajo porcentaje de instancias vinculadas a repositorios externos por lo que sería necesario incidir en este aspecto para aumentar la dimensión de enlazado.

Tabla 6.7: Resumen de los resultados obtenidos por dimensión para la calidad de los datos.

Dimensión	Criterio	Valor
Precisión	Validez sintáctica de los documentos RDF	1
	Validez sintáctica de los literales	0,9976

Sigue en la página siguiente.

Dimensión	Criterio	Valor
	Validez semántica de tripletas	1
Fiabilidad	Integridad a nivel de conjunto de datos	0,25
	Integridad a nivel de instancia	0
	Uso de valores vacíos y desconocidos	0
Consistencia	Consistencia de las restricciones de esquema durante la inserción de instancias	0
	Consistencia de las instancias con respecto a las restricciones de clase	1
	Consistencia de las instancias con respecto a las restricciones de relación	1
Relevancia	Crear un ranking de instancias	0
Integridad	Integridad de esquema	0,8
	Integridad de columna	0,8
	Integridad de población	0,625
Actualidad	Frecuencia de actualización	0,25
	Especificación del periodo de validez de las instancias	0
	Especificación de la fecha de modificación de las instancias	0
Facilidad de entendimiento	Descripción de los recursos	0,12
	Etiquetas en varios idiomas	0
	Serialización comprensible del RDF	1
	URIs autodescriptivos	1
Interoperabilidad	Evitar nodos en blanco y la cosificación del RDF	1
	Proveer diferentes formatos de serialización	1
	Uso de vocabulario externo	0,57

Sigue en la página siguiente.

Dimensión	Criterio	Valor
	Interoperabilidad del vocabulario propietario	1
Accesibilidad	Posibilidad de desreferenciar recursos	1
	Disponibilidad del conjunto de datos	1
	Disponibilidad de un punto de acceso SPARQL público	1
	Opción de descarga en RDF	1
	Negociación de contenido	0,5
	Enlazado del sitio HTML con la serialización RDF	0
	Suministro de metadatos del repositorio	1
Licencia	Suministro de información legible por máquina sobre la licencia	1
Enlazado	Enlazado mediante `owl:sameAs`	0,12
	Validez de los URIs externos	1

Como conclusión de este paso, decir que se ha realizado una evaluación cuantitativa de la calidad de los datos basándose en la propuesta de [Färber et al., 2018]. Los resultados obtenidos proporcionan una imagen completa de la calidad que ofrece este conjunto de datos (ver tabla 6.7). Cabe destacar los buenos resultados obtenidos en las dimensiones de *precisión*, que prueba datos correctos y confiables; en la *integridad*, que evidencia un repositorio con suficiente amplitud y alcance; la *interoperabilidad*, esencial para el intercambio y reutilización de los datos; la *accesibilidad*, cuyo resultado demuestra la disponibilidad y recuperación de los datos así como su descarga; y por último, la dimensión de *licencia*, ya que facilitar esta información resulta fundamental para promover la reutilización de datos. La información obtenida en este paso se puede utilizar como retroalimentación de un proceso de mejora de la calidad incidiendo en aquellos aspectos que

han obtenido un puntaje más bajo como es el caso del *enlazado*, *relevancia* y *fiabilidad*.

6.7 Explotación de los datos

Este es el último paso del modelo propuesto y cubre la explotación del conjunto de datos, dando solución tanto a usuarios no expertos (sin demasiados conocimientos sobre las TIC) como expertos. Además, facilita que las aplicaciones puedan conectarse automáticamente para consultar y recuperar la información contenida en el conjunto de datos, promoviendo así su reutilización.

Para aprovechar el potencial del conjunto de datos y evitar la complejidad asociada por ejemplo, al uso de SPARQL, RDF e incluso las ontologías, la propuesta utiliza CubeViz.js para proporcionar una visualización y explotación estadística de los datos. La aplicación proporciona un conjunto completo de funciones en formato cuadro de mando para seleccionar y visualizar observaciones[20] que ayuda en el proceso de toma de decisiones.

La figura 6.15 muestra un ejemplo de interfaz web para interactuar con los datos generados de acuerdo al vocabulario RDF Data Cube. Esta visualización facilita la consulta y comprensión de los datos y permite su manejo de forma sencilla. En esta figura aparece un menú superior que permite seleccionar el *dataset*, las medidas, atributos y dimensiones para su visualización.

Las figuras 6.16, 6.17 y 6.18 son los gráficos obtenidos con el cuadro de mando después de filtrar por la medida *Critical cleaning spots*, *Population* y *Per capita income* respectivamente, en las dimensiones de *Region* para una selección de barrios y *Time* seleccionando el periodo de tiempo a mostrar, en este ejemplo el mes de febrero de 2017.

[20] `https://smartdataua.github.io/rdfdatacube/`

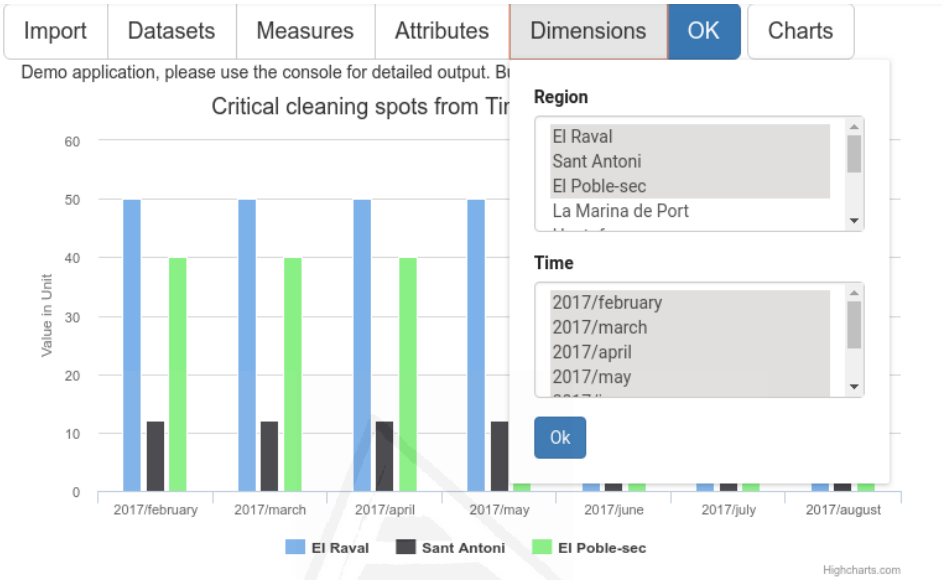


Figura 6.15: Representación gráfica de los puntos críticos de limpieza a partir de las dimensiones *Region* y *Time* (meses del 2017).

Teniendo en cuenta los gráficos generados con el filtrado descrito en el párrafo anterior, se observa que no parece existir una correlación entre el número de residentes y el número de puntos críticos de limpieza. Así como que la cantidad de puntos críticos de limpieza no cambia significativamente a lo largo el año (figura 6.15). Sin embargo, sí se observa la existencia de una relación entre el número de puntos críticos de limpieza y el nivel de ingresos (per cápita) en áreas como *El Raval* y *El Poble Sec* (figura 6.18)[21].

Además de esta visualización de los datos, se habilita un punto de

[21]Son resultados extraídos de la observación de las gráficas.

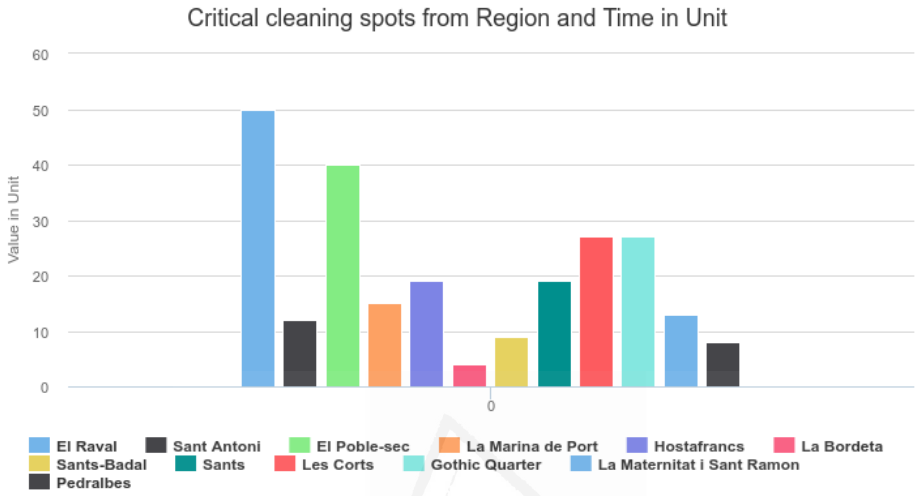


Figura 6.16: Representación gráfica de los puntos críticos de limpieza en las dimensiones *Region* (subconjunto de zonas seleccionadas) y *Time* (febrero de 2017).

acceso SPARQL para facilitar el acceso y la reutilización del conjunto de datos por usuarios expertos y permita la conexión automática de aplicaciones. Un punto de acceso SPARQL no solo facilita el acceso a los datos, sino que también permite las consultas federadas y que estas se ejecuten en otros interfaces con acceso SPARQL.

El enlazado con otras fuentes de datos permite agregar más información y mejorar el conjunto de datos final. En este sentido, Wikidata proporciona un conjunto completo de propiedades que pueden ser explotados, como subdivisiones administrativas, dimensiones, imágenes y proximidad geográfica. La figura 6.19 muestra un ejemplo de una consulta federada SPARQL que se utiliza para ejecutar consultas distribuidas en diferentes puntos SPARQL utilizando el comando

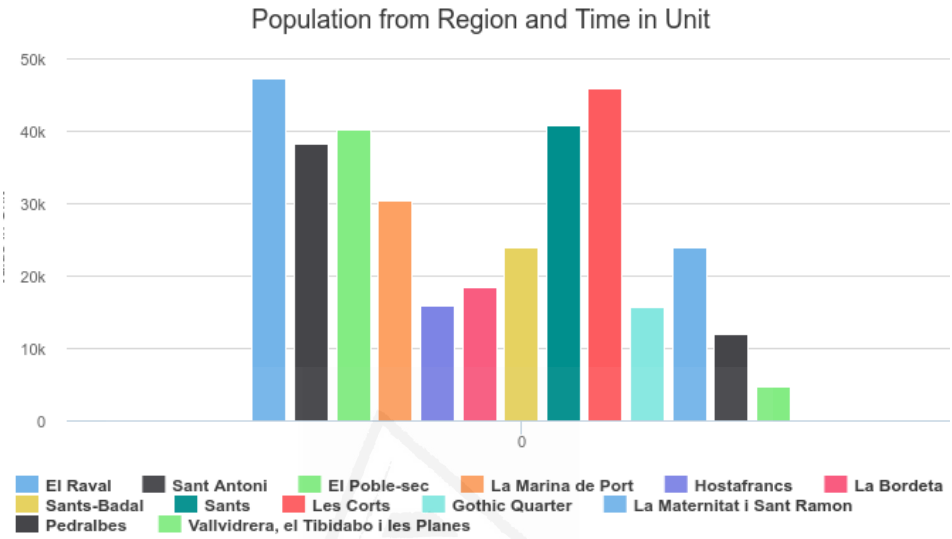


Figura 6.17: Representación gráfica de la población en las dimensiones de *Region* (zonas seleccionadas) y *Time* (febrero de 2017).

`SERVICE`. Esta consulta se ejecuta desde el punto SPARQL habilitado para esta experimentación (figura 6.20) y en su resultado combina datos del repositorio local con los de Wikidata: coordenadas geográficas `?coordinates`[22], área ocupada por una región `?area`[23] y otros identificadores externos adicionales como el de OpenStreetMap (OSM) `?osmid`[24]. Los resultados de esta consulta se muestran en la figura 6.21 y este sería un ejemplo de cómo los usuarios expertos pueden explotar este conjunto de datos.

[22] `https://www.wikidata.org/wiki/Property:P625`
[23] `https://www.wikidata.org/wiki/Property:P2046`
[24] `https://www.wikidata.org/wiki/Property:P402`

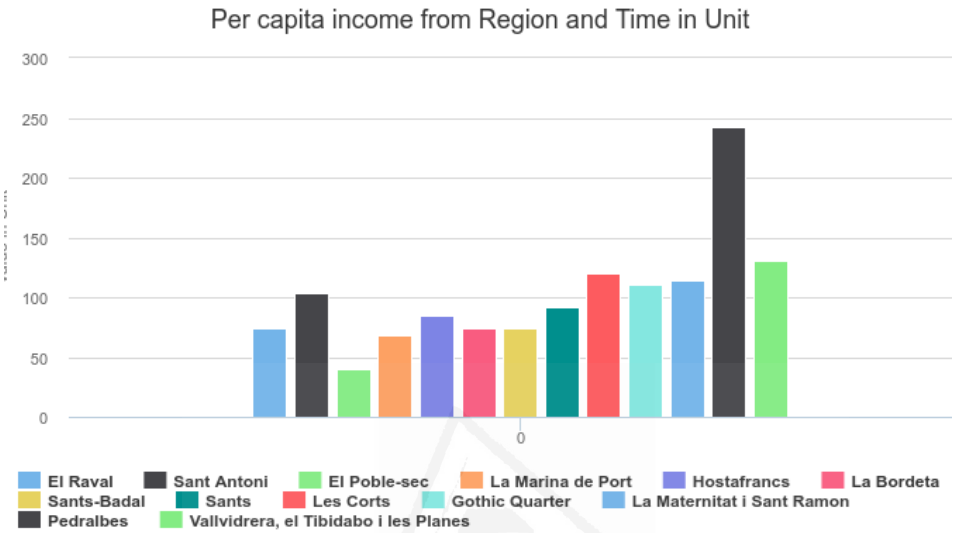


Figura 6.18: Representación gráfica de la renta per cápita en las dimensiones *Region* (zonas seleccionadas) y *Time* (febrero de 2017).

6.8 Valoración mediante criterio de expertos

En esta sección se describe el procedimiento aplicado para poder obtener la valoración del modelo refinado. En esta segunda iteración el modelo ha sido valorado por el mismo grupo, formado por 15 expertos, que realizó la valoración en la primera iteración (capítulo 5, sección 5.6).

Al igual que en la iteración anterior, se suministró información sobre el modelo y las fases que lo componen aplicado a un caso real visto en este capítulo, así como el objetivo principal que debía cumplir. Además, se realizaron entrevistas individuales con el fin de poder

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX ex: <http://example.com/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>

SELECT DISTINCT ?s ?area ?coordinates ?osmid
WHERE {
  ?s a ex:Region .
  ?s owl:sameAs ?wd .
  SERVICE <https://query.wikidata.org/sparql>
  {
    ?wd wdt:P625 ?coordinates .
    ?wd wdt:P2046 ?area .
    ?wd wdt:P402 ?osmid .
  }
}
```

Figura 6.19: Consulta federada SPARQL que recupera información adicional a las regiones como las coordenadas geográficas en Wikidata, el área de la ubicación y el identificador OpenStreetMap (OSM).

aclarar las dudas sobre los cambios aplicados en el refinamiento del modelo, como el paso de evaluación de la calidad de los datos y otros aspectos más técnicos. Los datos de valoración de los expertos se recogieron a través de una encuesta que refleja las mejoras introducidas en el modelo para que puedan ser valoradas (ver anexo H para consultar la encuesta).

Se utilizó de nuevo la escala de Likert para procesar los resultados de las encuestas y obtener el grado de aceptación por parte de los expertos para esta nueva propuesta del modelo. Se identificó la frecuencia de cada uno de los niveles de la escala de Likert y después se calculó el índice porcentual IP para obtener el grado de aceptación del grupo sobre cada uno de los indicadores en cuestión. En el anexo

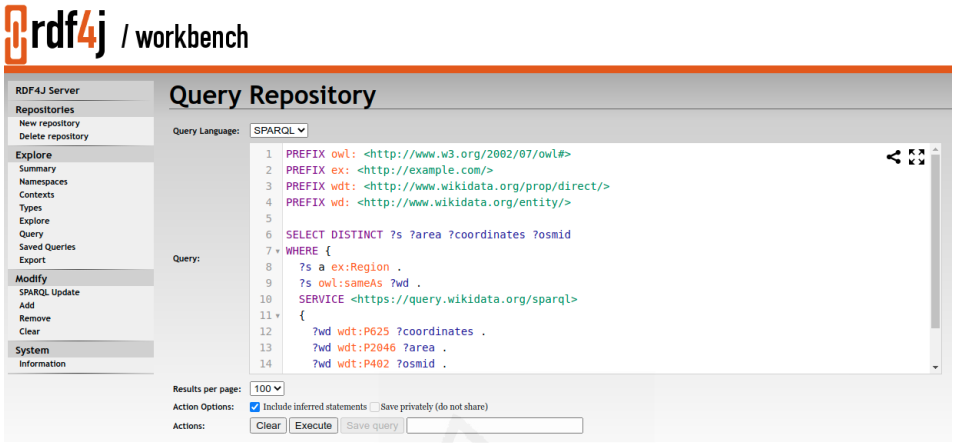
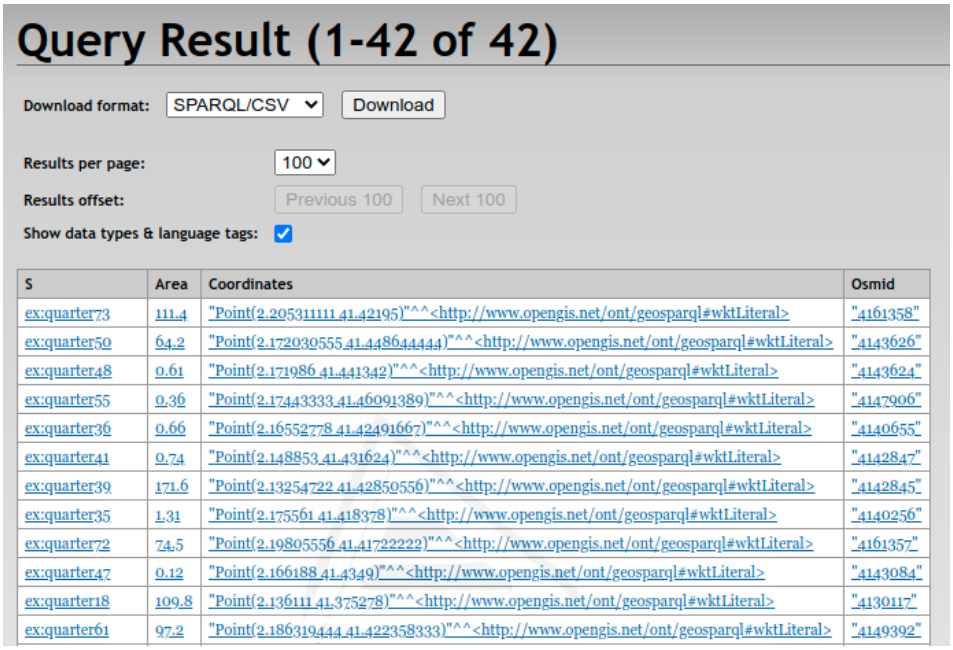


Figura 6.20: Editor SPARQL de RDF4J en el que se ejecuta una consulta federada que recupera información de Wikidata.

G se puede consultar los detalles sobre el proceso de cálculo del IP así como los indicadores valorados en esta segunda iteración.

El procesamiento de los datos realizado, evidencia que el modelo planteado, tanto en los principios fundamentales que debe cumplir como en las diferentes fases que lo componen, tienen una alta aceptación por parte de los expertos. La tabla G.1 recoge los valores obtenidos para cada uno de los niveles de la escala en cada uno de los indicadores de la encuesta. Y como se puede observar se ha obtenido un grado de aceptación de 88 % o más en todos los indicadores.

Los cuatro primeros indicadores referencian los principios fundamentales asociados al modelo (interoperabilidad, reutilización, calidad y enriquecimiento). En relación a los valores obtenidos en la iteración 1 (ver tabla E.1), se observa un incremento de aceptación en la reutilización y enriquecimiento de los datos. Pero el cambio más destacable es en el hecho de que el modelo refleje la calidad de los datos don-



Query Result (1-42 of 42)

Download format: SPARQL/CSV Download

Results per page: 100

Results offset: Previous 100 Next 100

Show data types & language tags:

S	Area	Coordinates	Osmid
ex:quarter73	111.4	"Point(2.205311111 41.42195)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4161358"
ex:quarter50	64.2	"Point(2.172030555 41.448644444)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4143626"
ex:quarter48	0.61	"Point(2.171986 41.441342)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4143624"
ex:quarter55	0.36	"Point(2.17443333 41.46091389)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4147906"
ex:quarter36	0.66	"Point(2.16552778 41.42491667)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4140655"
ex:quarter41	0.74	"Point(2.148853 41.431624)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4142847"
ex:quarter39	171.6	"Point(2.13254722 41.42850556)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4142845"
ex:quarter35	1.31	"Point(2.175561 41.418378)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4140256"
ex:quarter72	74.5	"Point(2.19805556 41.41722222)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4161357"
ex:quarter47	0.12	"Point(2.166188 41.4349)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4143084"
ex:quarter18	109.8	"Point(2.136111 41.375278)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4130117"
ex:quarter61	97.2	"Point(2.186319444 41.422358333)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>	"4149392"

Figura 6.21: Listado con los resultados obtenidos al ejecutar la consulta federada de la figura 6.19, en él se muestran los datos obtenidos de Wikidata: `Coordinates`, `Area` y `Osmid`.

de el grado de aceptación pasa de un 40 % al 90, 67 en esta segunda iteración,

La nueva propuesta del modelo contempla 6 pasos en vez de cuatro, y estos mantienen su valoración tanto para la fase de integración de datos como para el modelado y generación de RDF. Y sube en la publicación y sobre todo en la explotación.

Después de procesar los resultados obtenidos se puede decir que la mejora del modelo obtiene un grado de aceptación alto. Y que añadir

un paso para evaluar la calidad de los datos repercute directamente en los principios de reutilización y calidad así como en la publicación y explotación del contenido.

En lo que a la reutilización se refiere, ofrecer fiabilidad en los datos suministrados facilita que puedan ser utilizados en procesos de empresa con total confianza. Sin embargo sigue habiendo aspectos a mejorar en la evaluación de la calidad como marcar el origen del repositorio, comprobar la fiabilidad de la fuente de datos o marcar con metadatos los valores obtenidos de la evaluación.

7. Valoración del modelo mediante criterio de usuarios

> Sin datos, solo eres otra persona con una opinión.
>
> *Edwards Deming*

Los datos abiertos enlazados son, por su naturaleza, un recurso abierto, accesible por cualquier persona u organización que quiera consultarlos independientemente de su formación o perfil tecnológico, están disponibles para todos. Sin embargo, a día de hoy sigue existiendo una brecha digital en lo que respecta a las herramientas y experiencia de acceso a los datos.

El modelo descrito en el capítulo 4 no termina con la publicación de los datos abiertos enlazados sino que también facilita la explotación y reutilización ofreciendo un punto de acceso público SPARQL y un cuadro de mando que facilita la gestión y comprensión de los datos. Resulta fundamental realizar la valoración mediante el criterio de usuarios y analizar los resultados obtenidos.

En este capítulo se describe cómo se ha llevado a cabo dicha valoración en la que se pretende obtener el grado de satisfacción de los usuarios con los datos y su representación (paso de `Explotación`). Para realizar esta valoración se toma el cuadro de mando generado en la segunda iteración (sección 6.7) dado que corresponde a la aplicación del modelo ya refinado.

Para realizar la valoración del modelo se realiza una encuesta a un grupo de 20 usuarios[1] y se procesan los resultados obtenidos para extraer conclusiones sobre dicha valoración. Se ha diseñado una encuesta compuesta por cuatro bloques: en el primero se solicita al usuario información de carácter general; en el segundo se presentan tres preguntas de autoevaluación sobre la experiencia con los datos abiertos, hojas de cálculo y SPARQL; en el tercero se proponen 9 cuestiones a resolver con la ayuda de los datos suministrados en la prueba; y en el cuarto bloque, se pide valorar la experiencia de acuerdo a los aspectos de utilidad, calidad, eficiencia (obteniendo el grado de satisfacción de los usuarios con los datos e interfaz). La encuesta ha sido realizada con la herramienta de Google Form y se puede consultar en el anexo I.

Se realizan dos pruebas. En la primera se proporcionan los datos en bruto (formato de hoja de cálculo) y deben rellenar la encuesta ayudándose de estos datos. En la segunda, la información se facilita mediante un cuadro de mando generado por el modelo propuesto en este trabajo de tesis.

Se parte de la suposición de que los datos que se generan con la aplicación del modelo, son útiles para los usuarios y que facilitándoles un cuadro de mando mejora la comprensión y utilidad de los mismos. Para ello se comprueba si la puntuación obtenida en la primera prueba, difiere significativamente de la puntuación de la segunda (rellenando el mismo cuestionario) después de facilitarles un cuadro de mando para la gestión y visualización de los datos en un modelo multidimensional. Las pruebas se realizaron con dos meses de diferencia.

La tabla 7.1 muestra los resultados obtenidos en las dos pruebas. La columna `Hojas de calculo` representa el número de respuestas correctas obtenidas en la primera prueba (datos proporcionados en formato de hoja de cálculo) y `Cuadro de mando` las de la segunda.

[1]Muestra representativa de conveniencia formada por 20 usuarios con un perfil poco técnico y no experto en datos abiertos.

La última columna (d_j) muestra la diferencia entre la puntuación de ambas pruebas.

Tabla 7.1: Resultados de las encuestas realizadas por los usuarios.

Usuario	Hojas de cálculo	Cuadro de mando	d_j
1	4,00	8,00	-4
2	9,00	9,00	0
3	6,00	7,00	-1
4	7,00	7,00	0
5	7,00	6,00	1
6	5,00	6,00	-1
7	6,00	6,00	0
8	4,00	7,00	-3
9	3,00	7,00	-4
10	6,00	8,00	-2
11	7,00	8,00	-1
12	8,00	9,00	-1
13	7,00	9,00	-2
14	9,00	9,00	0
15	8,00	9,00	-1
16	5,00	7,00	-2
17	5,00	7,00	-2
18	8,00	8,00	0
19	4,00	5,00	-1
20	8,00	8,00	0

Se pretende averiguar si existen diferencias significativas entre ambas formas de presentar los datos a los usuarios. Para confirmarlo se realiza la prueba t para dos muestras relacionadas. Se utiliza un diseño apareado, adecuado para muestras que no son independientes. En este caso se parte de un diseño en el que el mismo grupo de usuarios son medidos en las dos pruebas.

Tabla 7.2: Resultados del test Shapiro-Wilk.

	Hojas de cálculo	Cuadro de mando
Test Shapiro-Wilk	0,9448891878128052	0,2960801124572754
p-value	0,9077096581459045	0,05765152722597122

Antes de aplicar esta prueba es necesario comprobar que los datos de las muestras se distribuyen de manera normal. Para ello se realiza el test de Shapiro-Wilk (SW) [SHAPIRO and WILK, 1965] para cada una de las muestras (ver tabla 7.2), de manera que si p-value $< 0,05$ los datos siguen una distribución normal. Como se puede observar ambos valores p-value son $< 0,05$[2], en consecuencia se puede afirmar con un nivel de confianza del 95 %, que no existen evidencias para rechazar la hipótesis de normalidad de los datos.

Se plantea la siguiente hipótesis nula H_0: *no existe una diferencia significativa entre los resultados obtenidos cuando se suministran los datos en bruto (`Hojas de cálculo`) y cuando se suministra un modelo multidimensional gestionado a través de un cuadro de mando (`Cuadro de mando`).* Siendo por tanto la hipótesis alternativa H_1: *existe una diferencia significativa en los resultados obtenidos cuando se suministran los datos en bruto y mediante un cuadro de mando.* Se realiza el siguiente contraste de hipótesis

$$H_0 \equiv \mu_d = 0$$
$$H_1 \equiv \mu_d \neq 0$$

Donde $H_0 \equiv \mu_1 = \mu_2$ y si se denomina $\mu_d = \mu_1 - \mu_2$ la hipótesis H_0 equivale a $H_0 \equiv \mu_d = 0$ y $H_1 \equiv \mu_1 \neq \mu_2$ sería equivalente a $H_1 \equiv \mu_d \neq 0$. Se fija un error $\alpha = 0,05$ para obtener un nivel de confianza del 95 %. Se aplica el estadístico t_0 según la fórmula 7.1 y

[2]El test de Shapiro-Wilk se ha realizado utilizando la librería *scipy.stats* de Python. Se ha calculado *shapiro_stat*, *shapiro_pvalor*.

suponiendo una normalidad demostrada anteriormente.

$$t_0 = \frac{\overline{d}}{\frac{s_d}{\sqrt{n}}} \tag{7.1}$$

Para este contraste se calculan los valores $\overline{d}$ y s_d^2 a partir de la fórmulas 7.2 y 7.3 respectivamente, donde n es el tamaño de la muestra y d_j la diferencia entre las muestras.

$$\overline{d} = \frac{1}{n-1}\sum_{j=1}^{n} d_j \tag{7.2}$$

$$s_d^2 = \frac{1}{n-1}\sum_{j=1}^{n}(d_j - d)^2 \tag{7.3}$$

A continuación se muestran los valores obtenidos:

$$\overline{d} = -1,20$$
$$s_d^2 = 1,85$$
$$s_d = 1,36$$
$$n = 20$$

Se aplica el estadístico t_0 y se compara con el valor de una distribución *T de Student* $t_{\frac{\alpha}{2},(n-1)}$ de manera que si resulta $|t_0| > t_{\frac{\alpha}{2},(n-1)}$ se rechazaría la hipótesis nula $\mu_d = 0$.

Una vez calculado $|t_0| = 3,94$, se compara con el valor de la distribución para $t_{\frac{\alpha}{2},(n-1)} = 2,093$ de manera que como se cumple $|t_0| = 3,94 > t_{\frac{0,05}{2},19} = 2,093$, se rechaza la hipótesis nula H_0. Por consiguiente, se concluye que existen diferencias significativas en los resultados obtenidos en función de las dos formas de presentar los datos a los usuarios, a un nivel de confianza del 95 %.

Tabla 7.3: Valores sobre la satisfacción del grupo de usuarios con la presentación de los datos en la primera prueba (`Hojas de cálculo`) según la escala de Likert.

	CD		ED		N-Si-No		DA		MA		
Aspectos	Núm.	%	Núm.	%	Núm.	%	Núm.	%	Núm.	%	IP
Satisfacción con la presentación	16	80,00	3	15,00	0	0,00	1	5,00	0	0,00	26,00
Utilidad de los datos	4	20,00	1	10,00	13	65,00	1	5,00	0	0,00	51,00
Evaluación de la calidad	0	0,00	0	0,00	2	10,00	3	15,00	15	75,00	93,00

Por otro lado, se desea obtener el grado de aceptación de los usuarios con los datos proporcionados (en ambas pruebas) y su representación. Para procesar los resultados obtenidos se ha utilizado la escala de Likert para obtener el grado de acuerdo o desacuerdo del grupo de usuarios con cada una de los aspectos planteados en la encuesta (ver tablas tabla 7.3 y 7.4). Se identifica el número y frecuencia de cada uno de los niveles de esta escala y después se calcula el índice porcentual IP, que informa sobre el valor de aceptación del grupo sobre el aspecto en cuestión.

Para la primera encuesta (tabla 7.3) se ha obtenido un índice porcentual de $26,00\,\%$, lo que representa un valor muy bajo de aceptación por el grupo de usuarios. Sin embargo, para la segunda prueba (tabla 7.4) se obtiene un grado de aceptación alto, del $95,00\,\%$. Con respecto a si les han resultado útiles los datos, la primera prueba ha obtenido un índice porcentual de $51,00\,\%$ frente a $91,00\,\%$ de la segunda, en la que hay mayor grado de aceptación con respecto a la utilidad de los datos. Y por último, se les pide que valoren la importancia de la calidad en los datos y en ambas pruebas se han obtenido $93,00\,\%$.

Los resultados obtenidos sobre el grado de aceptación por parte de los usuarios, confirman la conclusión extraída anteriormente en la que se rechazaba la hipótesis nula H_0 y en la que se evidencian diferencias significativas en los resultados obtenidos en los dos supuestos. Siendo el grado de aceptación de la segunda prueba (cuadro de mando) $95,00\,\%$, se puede concluir que es más eficiente suministrar una representación

Tabla 7.4: Valores sobre la satisfacción del grupo de usuarios con la presentación de los datos en la segunda prueba (`Cuadro de mando`), según la escala de Likert.

	CD		ED		N-Si-No		DA		MA		
Items	Núm.	%	Núm.	%	Núm.	%	Núm.	%	Núm.	%	IP
Satisfacción con la presentación	0	0,00	0	0,00	1	5,00	3	15,00	16	80,00	95,00
Utilidad de los datos	0	0,00	0	0,00	2	10,00	5	25,00	13	65,00	91,00
Evaluación de la calidad	0	0,00	0	0,00	2	10,00	3	15,00	15	75,00	93,00

de los datos en un modelo multidimensional y gestionado a través de un cuadro de mando.

En consecuencia, se puede afirmar que se generan datos útiles para los usuarios no expertos ayudándoles en la resolución o toma de decisiones. Por lo que se apoyaría la hipótesis planteada al inicio de este documento de tesis (capítulo 1) aplicando un método de transformación a un conjunto de fuentes diversas de datos utilizando un modelo de enfoque multidimensional se pueden generar datos útiles para usuarios expertos y no expertos.

La muestra recoge usuarios de edades comprendidas entre 28 y 68 años, resaltando que 50 % son mujeres, 45 % hombres y una persona que prefirió no dar esta información. Además, se realizaron tres preguntas de autoevaluación sobre sus conocimientos en datos abiertos, hojas de cálculo y SPARQL. El 85 % de los usuarios encuestados no tiene conocimientos sobre datos abiertos y el 95 % reveló no conocer SPARQL. Sin embargo, en lo que respecta al conocimiento sobre hojas de cálculo, el 46 % afirma conocer y trabajar en ocasiones con hojas de cálculo.

Por último, si se observan los tiempos invertidos por los usuarios en rellenar las encuestas (ver figuras 7.1 y 7.2), se detectan tiempos más elevados en la primera encuesta, donde el $77,8\,\%$ de los usuarios tardo más de 5 minutos en contestar a las cuestiones, el $16,7\,\%$ lo hizo en aproximadamente 4 minutos y solo el 5,6 % lo hizo en 3 minutos. Frente a los datos recogidos en el segundo supuesto donde los tiempos

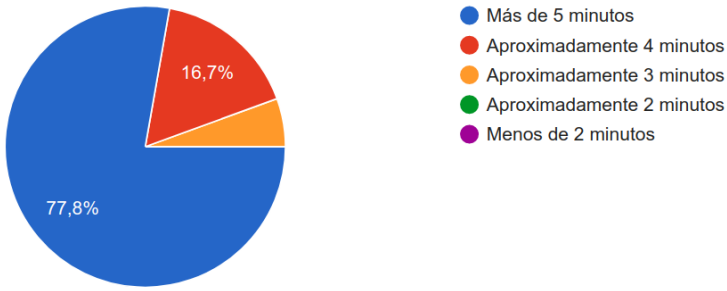


Figura 7.1: Distribución de los tiempos aproximados que los usuarios tardaron en realizar la primera encuesta.

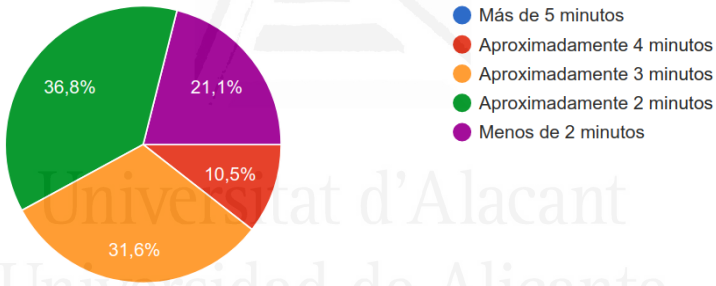


Figura 7.2: Distribución de los tiempos aproximados que los usuarios tardaron en realizar la segunda encuesta.

registrados son mucho más cortos, el $57,9\,\%$ lo hizo en 2 minutos o menos, el $31,6\,\%$ en unos 3 minutos y el $10,5\,\%$ en 4 minutos.

8. Conclusiones

> Nadie sabe el potencial que encierra este poderoso sistema. Algún día podrá llegar a ejecutar música, componer sinfonías y complejos diseños gráficos.
>
> *Ada Lovelace*

8.1 Aportaciones fundamentales de esta tesis

Se ha cubierto el objetivo general planteado en esta tesis: **Definir un modelo que, a partir de un conjunto de datos obtenidos de diversas fuentes, facilite la publicación, enriquecimiento y validación de LOD, generando información útil y de calidad orientada a usuarios expertos y no expertos**, con las siguientes aportaciones:

1. **Definir un modelo que, a partir de un conjunto de datos obtenidos de diversas fuentes, facilite la publicación.** La aportación más relevante de este trabajo es la propuesta de un modelo para la publicación de datos abiertos enlazados con un enfoque multidimensional basado en RDF, que garantiza la confiabilidad en los datos, mejorando la explotación y promoviendo

la su reutilización efectiva. La publicación del conjunto de datos basado en el vocabulario RDF Data Cube, sigue los principios planteados en el esquema de desarrollo de 5 estrellas propuesto por Tim Berners-Lee y promovido por el W3C.

Además, se ha realizado la valoración del modelo mediante criterio de expertos que con su opinión han contribuido al refinamiento del mismo en dos iteraciones. Se ha obtenido una alta aceptación tanto en los principios fundamentales que debe cumplir como en las diferentes fases que lo componen.

2. **Enriquecimiento**. Se han utilizado repositorios externos como Wikidata y GeoNames, como base de conocimiento para el enriquecimiento del conjunto de datos original.

3. **Validación de LOD**. Se ha incorporado una metodología para la evaluación de la calidad del conjunto de datos basado en el vocabulario RDF Data Cube. Se ha adaptado a las especificaciones de los repositorios de cubos de datos ya que su planteamiento inicial está orientado a KG.

4. **Generando información útil y de calidad orientada a usuarios expertos y no expertos**. La explotación del conjunto de datos está orientada tanto a usuarios expertos como no expertos, proporcionando un cuadro de mando que facilita a los usuarios utilizar y comprender de manera efectiva los datos generados de acuerdo al vocabulario RDF Data Cube. Además, se suministra un punto de acceso SPARQL público para usuarios expertos y para que las aplicaciones puedan conectar, consultar y recuperar la información contenida en el repositorio, así como ejecutar consultas federadas que permiten obtener otros atributos no contenidos en el repositorio original.

> Esta aportación se ha valorado mediante criterio de usuarios que ilustra cómo los datos generados y publicados a través de este modelo, en contexto y forma, son útiles para los usuarios. Y que las pruebas realizadas ofrecen un grado de confianza de 97, 33 %.

Cada vez más organizaciones públicas y privadas publican sus datos en abierto, cada vez se generan más datos originados por dispositivos, sensores o redes sociales. Pero, pese a la existencia de esta ingente cantidad de datos disponibles en Internet, resulta complejo reutilizarlos (acceder y explotar estos datos). Además, no siempre se tiene la certeza de acceder a datos confiables de fuentes contrastadas, disuadiendo a organizaciones y usuarios de contribuir y reutilizar esos datos.

Por tanto, resulta fundamental intensificar las acciones para garantizar el valor de los datos, evaluando y mejorando la calidad, enlazando con otros repositorios de calidad contrastada como Wikidata o GeoNames, y seleccionando una representación adecuada de la información que facilite su explotación.

La aportación de este trabajo de tesis es la definición de un modelo con un enfoque multidimensional basado en RDF para publicar y explotar LOD. Partiendo de un conjunto de datos recolectados de diversas fuentes, se genera información útil para los usuarios utilizando estándares de Web semántica publicados por el W3C. La principal motivación de esta investigación se basa en aumentar el valor de los datos, enriqueciéndolos semánticamente, evaluando la calidad, y proporcionando los mecanismos que faciliten su explotación (por ejemplo mediante cuadro de mando y un punto de acceso público SPARQL). Publicar no es suficiente, sino que es necesario promover su reutilización efectiva e innovadora.

LOD proporciona un enfoque innovador para la publicación y reutilización de los datos. Facilita la reutilización y el enriquecimiento mediante la conexión con fuentes externas, permitiendo la extensión

del modelo de datos. A través del uso de puntos de acceso SPARQL se realiza la integración de diferentes conjuntos de datos. Las consultas federadas permiten la ejecución de consultas distribuidas en diferentes puntos de acceso SPARQL. Además, permiten la inclusión de nuevos atributos en el conjunto de datos original (por ejemplo se han enriquecido los datos de localizaciones utilizadas en los dos casos de uso, obteniendo información más detallada sobre las zonas de suministro de agua visto en el capítulo 5 o de los barrios de Barcelona en el capítulo 6). A partir de esta información más completa, se pueden extraer nuevos indicadores, así como nuevas correlaciones entre indicadores que mejoran el proceso de toma de decisiones.

En los artículos [Escobar Esteban et al., 2020b,a] se describe la aplicación práctica del modelo presentado en este trabajo de tesis a dos casos reales (empresa de suministro de agua y la plataforma *Open Data BCN*, respectivamente).

Adicionalmente, se ha realizado la valoración del modelo en lo relativo a los principios que lo fundamentan y al diseño de sus fases. Esta valoración se ha realizado mediante criterio de expertos y los resultados muestran un grado de aceptación alto, dando por válido el desarrollo de este modelo y avalado por el criterio de los expertos. Por otro lado, se ha valorado mediante criterio de usuarios, la utilidad de los datos generados para los usuarios y los resultados evidencian un alto grado de aceptación frente a los datos en bruto.

La incorporación de una metodología para la evaluación de la calidad de los datos ha supuesto una mejora cualitativa dando un aspecto más formal y confiable a la publicación de LOD. Para promover el uso de LOD y fomentar su reutilización es necesario evaluar la calidad asegurando que los datos publicados son útiles y confiables para los usuarios. La reutilización de los datos con calidad contrastada incentiva su reutilización y por consiguiente, el desarrollo económico. Esta

metodología ha sido aplicada con éxito en [Candela et al., 2020][1]. En este estudio se realiza la evaluación de la calidad y posterior comparativa entre cuatro repositorios LOD de bibliotecas digitales, ofreciendo una visión general de la calidad de cada uno de los repositorios analizados.

8.2 Trabajo futuro

Se ha puesto de manifiesto la importancia que tiene la calidad de los datos para una reutilización eficiente. El último informe de 2019 sobre la madurez de los datos [European Data Portal, 2019] pone especial atención en la calidad, su reutilización e impacto.

Encontrar un método que pueda comprobar la calidad de los datos de forma automática, es costoso. Por ello es necesario seguir trabajando en esta línea. Por ejemplo, los resultados obtenidos en la fase de calidad de los datos, servirán de retroalimentación para identificar aquellas características que mejoren la publicación del conjunto de datos, como la incorporación de información sobre la procedencia o clasificación de los datos publicados. Se estudiará una mayor generalización y automatización de los procedimientos de evaluación. Del mismo modo, se analizará cómo añadir más puntos de control durante el proceso, para que expertos en el área puedan validar cada fase desde la generación de los datos hasta la generación del cuadro de mando. Además, sería recomendable publicar los resultados obtenidos sobre la calidad del conjunto de datos como LOD.

La calidad es crucial para promover la reutilización de los datos. Sería interesante estudiar cómo técnicas de Inteligencia Artificial (IA) podrían mejorar los procesos de evaluación de la calidad de los datos.

[1] Aceptado, pendiente de publicación en Journal of Information Science 2020-05-12

Al ser este proceso muy heurístico es susceptible de ser sustituido por un modelo neuronal con entrenamiento supervisado.

En el capítulo 4 se ha presentado una modelización desde el punto de vista de la gestión de los procesos de negocio, (utilizando BPMN), pero sería recomendable tener un enfoque *Model-Driven Architecture* (MDA) que permita diseñar la arquitectura del sistema TIC especificando los tres niveles de modelización *Platform-Independent Model* (PIM), *Platform Definition Model* (PDM) y *Platform-Specific Models* (PSM).

Finalmente, en la valoración realizada por los usuarios se muestra una clara orientación hacia presentaciones más visuales, sencillas de entender que permiten a usuarios no expertos identificar los datos clave para la consecución de su objetivo, por ello, es necesario realizar la extensión del modelo a representaciones temporales y geográficas en busca de una mejora en la explotación de LOD por parte de los usuarios. Por último, como trabajo futuro se plantea ofrecer a los usuarios una interfaz más amigable y sencilla, como un asistente virtual que sea capaz de dar respuesta a los usuarios consultando diferentes repositorios de LOD.

9. Difusión de la investigación

> El que lee mucho y anda mucho, ve mucho y sabe mucho.
>
> *Miguel de Cervantes*

El trabajo realizado en esta tesis ha dado lugar a las siguientes publicaciones en revistas indexadas en JCR:

1. Pilar Escobar, Gustavo Candela, Juan Trujillo, Manuel Marco-Such, Jesús Peral. Adding value to Linked Open Data using a multidimensional model approach based on the RDF Data Cube vocabulary. (2020). Computer Standards & Interfaces, 68, 103378. doi:10.1016/j.csi.2019.103378. F.I. (2018): 2.441. Computer Science, Software Engineering: Q2,D3 (29/107).

2. Pilar Escobar, María del Mar Roldán-García, Jesús Peral, Gustavo Candela, José García-Nieto. An Ontology-Based Framework for Publishing and Exploiting Linked Open Data: A Use Case on Water Resources Management. (2020). Applied Sciences, 10(3), 779. Section: Computing and Artificial Intelligence. doi:10.3390/app10030779. F.I. (2018): 2.217. Physics, Applied: Q2 (67/148).

3. Gustavo Candela, Pilar Escobar, Rafael C. Carrasco, Manuel Marco-Such. Evaluating the quality of linked open data in digital libraries. Journal of Information Science. (2020). F.I. (2018):

2.327. Computer Science, Information Systems: Q2 (74/155). Aceptado pendiente de publicación en Journal of Information Science 2020-05-12.

Junto con el equipo de tecnología e investigación de la Biblioteca Virtual Miguel de Cervantes se han realizado las siguientes actividades:

1. Difusión y fomento de tecnologías y datos abiertos enlazados a través de ejemplos prácticos en `data.cervantesvirtual.com`[1]. Además, la Biblioteca Virtual Miguel de Cervantes pertenece a la *International Community GLAM Labs*[2] que tiene objetivo la reutilización de las colecciones de forma innovadora y creativa. Y como parte de esta comunidad se ofrecen herramientas que explotan y reutilizan las colecciones digitales, así como conjuntos de datos con licencias abiertas y ejemplos de utilización [S-C. et al., 2019].

2. El cuadro de mando para la visualización de los datos (basado en RDF Data Cube) está disponible en Github[3] para su descarga y ejecución.

3. Inclusión en la plataforma de Open Data Barcelona el cuadro de mando para la visualización y gestión de datos abiertos enlazados basada en los datos abiertos publicados sobre las visitas y estado de los puntos críticos de la ciudad de Barcelona (consultar la figura 9.1). Estos datos han sido procesados y enriquecidos para publicarlos como LOD.

[1] Lab `http://data.cervantesvirtual.com/blog/`
[2] `https://glamlabs.io/`
[3] `https://smartdataua.github.io/rdfdatacube/`

Figura 9.1: Historias de uso realizadas con los datos publicados en *Open Data Barcelona*.

4. Esta investigación está enmarcada en el proyecto ECLIPSE- Enhancing Dala Quality and Security for Improving Business Processes and Strategic Decisions in Cyber Physical Systems.

5. Participación en la Red de investigación de Calidad y Sostenibilidad del Software (CALESI) en la que se presentó la metodología de evaluación de la calidad LOD. Valencia 2019.

6. Participación en la Red estratégica para la promoción de las infraestructuras de tecnologías del lenguaje en eHumanidades y ciencias sociales (INTELE) con propuestas para la explotación de contenido.

7. Además, el trabajo desarrollado con los datos abiertos en la plataforma `data.cervantesvirtual.com` fue galardonado con el segundo premio de los Premios Aporta organizado por red.es y el Ministerio de Energía, Turismo y Agenda Digital.[4] cuyo objetivo principal era el reconocimiento y difusión de proyectos innovadores desarrollados con datos públicos.

8. Gracias a este trabajo se han establecido colaboraciones con otras instituciones, en particular con el grupo *Ada Byron Research Building* de la Universidad de Málaga.

9. Fomentar la reutilización de los datos para crear nuevas aplicaciones, por ejemplo para realizar Trabajos Final de Grado. En el anexo J se pueden consultar algunos de eventos organizados para la promoción del uso de datos abiertos y los trabajos dirigidos por la autora de esta tesis dentro de este contexto.

Otras publicaciones (en revistas indexadas en JCR) vinculadas y que forman parte de los antecedentes de esta investigación:

[4] `http://datos.gob.es/es/noticia/el-buscador-linknovate-y-la-biblioteca-virtual-cervantes-premios-aporta-2017`

1. Gustavo Candela, Pilar Escobar, Rafael C. Carrasco, Manuel Marco-Such. A linked open data framework to enhance the discoverability and impact of culture heritage. Journal of Information Science, 45(6), 756-766. (2019). doi:10.1177/0165551518812658. F.I. (2018): 2.327. Computer Science, Information Systems: Q2 (74/155).

2. Gustavo Candela, Pilar Escobar, Rafael C. Carrasco, Manuel Marco-Such. Migration of a library catalogue into RDA linked open data. Semantic Web, 9(4), 481-491. (2018). doi:10.3233/SW-170274. F.I. (2018): 3.524. Computer Science, Information Systems: Q1 (35/155).

3. Linguistically-Enhanced Search over an Open Diachronic Corpus. ECIR. 2015. Clase 3 (The GII-GRIN Conference Rating 2015).

Otras publicaciones y presentaciones en las que se ha difundido la aplicación de estas metodologías muchas de ellas centradas en el dominio de las bibliotecas digitales.

1. María Dolores Sáez, Pilar Escobar, Gustavo Candela, Manuel Marco-Such and Mahendra Mahey. Publishing and reusing Linked Open Data in Galleries, Libraries, Archives and Museums (GLAMs): opportunities and challenges. El museo para todas las personas. Madrid, 2019.

2. Gustavo Candela, Pilar Escobar, Borja Navarro-Colorado. In search of Poetic Rhythm: Poetry retrieval through text and metre. En Actas de 2nd International Conference on Digital Access to Textual Cultural Heritage, DATeCH 2017, pp. 53-57. (2017).

3. Gustavo Candela, Pilar Escobar, Manuel Marco-Such. Semantic Enrichment on Cultural Heritage collections: A case study using

geographic information. En Actas de 2nd International Conference on Digital Access to Textual Cultural Heritage, DATeCH 2017, pp. 169-174. (2017).

4. Rafael C. Carrasco, Isabel Martínez-Sempere, Enrique Mollá-Gandía, Felipe Sánchez-Martínez, Gustavo Candela, Pilar Escobar. Linguistically-Enhanced Search over an Open Diachronic Corpus. En Actas de European Conference on Information Retrieval, ECIR 2015, pp. 801-804. Springer, Cham. (2015). doi:10.1007/978-3-319-16354-3_89. Clase 3 (The GII-GRIN Conference Rating 2015).

5. Gustavo Candela, Pilar Escobar, Manuel Marco-Such, Rafael C. Carrasco. Transformation of a Library Catalogue into RDA Linked Open Data. En Actas de 19th International Conference on Theory and Practice of Digital Libraries TPDL 2015, pp. 321-325. Springer, Cham. (2015). doi:10.1007/978-3-319-24592-8_26. Clase 3 (The GII-GRIN Conference Rating 2017)/ Work in progress (The GII-GRIN Conference Rating 2015).

6. Corpus diacrónico de la Biblioteca Virtual Miguel de Cervantes. Simposi Internacional Canvi lexicosemàntic, cultura i lingüística de corpus. XII Simposi Internacional Noves Tendències en I+D+i en Literatura, Llengua, Ensenyament i TIC. De la Innovació al Cànon. Alicante, 2018.

Bibliografía

Alberto Abella and De-Pablos-Heredero Carmen Ortiz-de Urbina-Criado, Marta. Meloda, a metric to assess open data reuse. http://www.elprofesionaldelainformacion.com/contenidos/2014/nov/04.pdf, 2014. [Online; accessed 04-September-2019].

Konrad Abicht, Georges Alkhouri, Natanael Arndt, Roy Meissner, and Michael Martin. Cubeviz.js: A lightweight framework for discovering and visualizing RDF data cubes. In Maximilian Eibl and Martin Gaedke, editors, *47. Jahrestagung der Gesellschaft für Informatik, Informatik 2017, Chemnitz, Germany, September 25-29, 2017*, volume P-275 of *LNI*, pages 1915–1921. GI, 2017. doi: 10.18420/in2017_191. URL https://doi.org/10.18420/in2017_191.

Elise Acheson, Stefano De Sabbata, and Ross S. Purves. A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64:309–320, 2017. doi: 10.1016/j.compenvurbsys.2017.03.007. URL https://doi.org/10.1016/j.compenvurbsys.2017.03.007.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web*

Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007., volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2007. doi: 10.1007/978-3-540-76298-0_52. URL https://doi.org/10.1007/978-3-540-76298-0_52.

Srividya K. Bansal. Towards a semantic extract-transform-load (ETL) framework for big data integration. In *2014 IEEE International Congress on Big Data, Anchorage, AK, USA, June 27 - July 2, 2014*, pages 522–529. IEEE Computer Society, 2014. doi: 10.1109/BigData.Congress.2014.82. URL https://doi.org/10.1109/BigData.Congress.2014.82.

Fernando Benitez-Paez, Auriol Degbelo, Sergio Trilles, and Joaquín Huerta. Roadblocks hindering the reuse of open geodata in colombia and spain: A data user's perspective. *ISPRS Int. J. Geo-Information*, 7(1):6, 2018. doi: 10.3390/ijgi7010006. URL https://doi.org/10.3390/ijgi7010006.

David Wood Bernadette Hyland. The joy of data - cookbook for publishing linked government data on the web. http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook, 2011.

Devis Bianchini, Valeria De Antonellis, Massimiliano Garda, and Michele Melchiori. Exploiting smart city ontology and citizens' profiles for urban data exploration. In Hervé Panetto, Christophe Debruyne, Henderik A. Proper, Claudio Agostino Ardagna, Dumitru Roman, and Robert Meersman, editors, *On the Move to Meaningful Internet Systems. OTM 2018 Conferences - Confederated International Conferences: CoopIS, C&TC, and ODBASE 2018, Valletta, Malta, October 22-26, 2018, Proceedings, Part I*, volume 11229 of *Lecture Notes in Computer Science*, pages 372–389. Springer, 2018. doi: 10.1007/978-3-030-02610-3_21. URL https://doi.org/10.1007/978-3-030-02610-3_21.

Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009. URL https://eprints.soton.ac.uk/271285/.

Ana Brandusescu and Danny Lämmerhirt. Open Data Charter Measurement Guide. https://drive.google.com/file/d/1yNOPMP1IrO68l4Swg16zqD8aTWDgFml-/view, 2018. [Online; April-2019].

Julio Cabrero Almenara and Julio Barroso Osuna. The use of expert judgment for assessing ict: the coefficient of expert competence. *Bordón, Revista de Pedagogía*, 65, 2013. [Online; accessed June-2019].

Gustavo Candela, Pilar Escobar, Rafael C Carrasco, and Manuel Marco-Such. A linked open data framework to enhance the discoverability and impact of culture heritage. *Journal of Information Science*, 0(0):0165551518812658, 2019. doi: 10.1177/0165551518812658. URL https://doi.org/10.1177/0165551518812658.

Gustavo Candela, Pilar Escobar, Rafael Carrasco, and Manuel Marco-Such. Evaluating the quality of linked open data in digital libraries, 2020.

María Hallo Carrasco, Sergio Luján-Mora, and Alejandro Maté. Evaluating open access journals using semantic web technologies and scorecards. *J. Information Science*, 43(1):3–16, 2017. doi: 10.1177/0165551515624353. URL https://doi.org/10.1177/0165551515624353.

Simona Colucci, Francesco M. Donini, and Eugenio Di Sciascio. Reasoning over RDF knowledge bases: Where we are. In Floriana Es-

posito, Roberto Basili, Stefano Ferilli, and Francesca A. Lisi, editors, *AI*IA 2017 Advances in Artificial Intelligence - XVIth International Conference of the Italian Association for Artificial Intelligence, Bari, Italy, November 14-17, 2017, Proceedings*, volume 10640 of *Lecture Notes in Computer Science*, pages 243–255. Springer, 2017. doi: 10.1007/978-3-319-70169-1_18. URL https://doi.org/10.1007/978-3-319-70169-1_18.

Edward Curry, Viktoriya Degeler, Eoghan Clifford, Daniel Coakley, Andrea Costa, Schalk-Jan Van Andel, Nick Van De Giesen, Christos Kouroupetroglou, Thomas Messervey, Jan Mink, and Sander Smit. Linked Water Data For Water Information Management. In *11th International Conference on Hydroinformatics (HIC 2014)*, New York, New York, USA, 2014.

Richard Cyganiak and Dave Reynolds. The RDF Data Cube Vocabulary. https://www.w3.org/TR/vocab-data-cube/, 2014. [Online; accessed 8-April-2018].

Blodgett David, Read Emily, Lucido Jessica, Slawecki Tad, and Young Dwane. An analysis of water data systems to inform the open water data initiative. *Journal of the American Water Resources Association*, 52:845–858, 01 2016. doi: 10.1111/1752-1688.12417. URL http://pubs.er.usgs.gov/publication/70170185.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Haiwei Dong, Gobindbir Singh, Aarti Attri, and Abdulmotaleb El-Saddik. Open data-set of seven canadian cities. *IEEE Access*, 5: 529–543, 2017. doi: 10.1109/ACCESS.2016.2645658. URL https://doi.org/10.1109/ACCESS.2016.2645658.

María Pilar Escobar Esteban, gustavo Candela Romero, Juan Trujillo, Manolo Marco-Such, and Jesús Peral. Adding value to linked open data using a multidimensional model approach based on the RDF data cube vocabulary. *Comput. Stand. Interfaces*, 68, 2020a. doi: 10.1016/j.csi.2019.103378. URL https://doi.org/10.1016/j.csi.2019.103378.

María Pilar Escobar Esteban, María del Mar Roldán-García, Jesús Peral Cortés, Gustavo Candela Romero, and José García-Nieto. An ontology-based framework for publishing and exploiting linked open data: A use case on water resources management. *Appl. Sci.*, 10 (779), 2020b. URL https://www.mdpi.com/2076-3417/10/3/779.

Lorena Etcheverry and Alejandro A. Vaisman. QB4OLAP: A vocabulary for OLAP cubes on the semantic web. In Juan F. Sequeda, Andreas Harth, and Olaf Hartig, editors, *Proceedings of the Third International Workshop on Consuming Linked Data, COLD 2012, Boston, MA, USA, November 12, 2012*, volume 905 of *CEUR Workshop Proceedings*, pages–. CEUR-WS.org, 2012. URL http://ceur-ws.org/Vol-905/EtcheverryAndVaisman_COLD2012.pdf.

Lorena Etcheverry, Silvia A. Gómez, and Alejandro A. Vaisman. Modeling and querying data cubes on the semantic web. *CoRR*, abs/1512.06080, 2015. URL http://arxiv.org/abs/1512.06080.

European Data Portal. Open Data in a nutshell. https://www.europeandataportal.eu/en/providing-data/goldbook/open-data-nutshell, 2015. [Online; accessed 5-April-2018].

European Data Portal. Open data maturity report 2019. https://www.europeandataportal.eu/sites/default/files/open_data_maturity_report_2019.pdf, 2019. [Online; accessed 13-April-2019].

Angélica María Fabila Echauri, Hiroe Minami, and Manuel Jesús Izquierdo Sandoval. La escala de likert en la evaluación docente: acercamiento a sus características y principios metodológicos. http://revistas.ujat.mx/index.php/perspectivas/article/viewFile/589/494, 2013. [Online; January-2019].

FAIR Data Maturity Model WG. Fair data maturity mode lspecification and guidelines 2020. https://www.rd-alliance.org/groups/fair-data-maturity-model-wg, 2020.

Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. Linked data quality of dbpedia, freebase, opencyc, wikidata, and YAGO. *Semantic Web*, 9(1):77–129, 2018. doi: 10.3233/SW-170275. URL https://doi.org/10.3233/SW-170275.

David Faye, Olivier Curé, and G Blin. A survey of rdf storage approaches. *ARIMA Journal*, 15:11–35, 01 2012.

Erwin Folmer, Wouter Beek, Laurens Rietveld, Stanislav Ronzhin, Rutger Geerling, and Davey den Haan. Enhancing the usefulness of open governmental data with linked data viewing techniques. In *52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019*, pages 1–10, 2019. URL http://hdl.handle.net/10125/59728.

Luis Galárraga, Kim Ahlstrøm Meyn Mathiassen, and Katja Hose. Qboairbase: The european air quality database as an RDF cube. In *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23rd - to - 25th, 2017.*, 2017. URL http://ceur-ws.org/Vol-1963/paper507.pdf.

Marcos André Gonçalves, Bárbara Lagoeiro Moreira, Edward A. Fox, and Layne T. Watson. "what is a good digital library? A quality

model for digital libraries. *Inf. Process. Manage.*, 43(5):1416–1437, 2007. doi: 10.1016/j.ipm.2006.11.010. URL https://doi.org/10.1016/j.ipm.2006.11.010.

Nicola Guarino et al. Formal ontology and information systems. In *Proceedings of FOIS*, volume 98, pages 81–97. 1998.

Steve Harris, Andy Seaborne, and Eric Prud'hommeaux. Sparql 1.1 query language. *W3C recommendation*, 21(10), 2013.

Jan Hauke and Tomasz Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93, 2011.

Swati Hira and P.S. Deshpande. Data analysis using multidimensional modeling, statistical analysis and data mining on agriculture parameters. *Procedia Computer Science*, 54:431 – 439, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.06.050. URL http://www.sciencedirect.com/science/article/pii/S1877050915013745. Eleventh International Conference on Communication Networks, ICCN 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Data Mining and Warehousing, ICDMW 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Image and Signal Processing, ICISP 2015, August 21-23, 2015, Bangalore, India.

Chia-Chien Hsu and Brian A. Sandford. The delphi technique: Making sense of consensus. *Practical Assessment, Research, and Evaluation*, 12(10), 2007. doi: 10.7275/pdz9-th90.

INE. Estadística sobre el suministro y saneamiento del agua, año 2016. notas de prensa. instituto nacional de estadística. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&

cid=1254736176834&menu=ultiDatos&idp=1254735976602, 2018. [Online; accessed 3-December-2019].

Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos A. Tarabanis. Linked open government data analytics. In Maria Wimmer, Marijn Janssen, and Hans Jochen Scholl, editors, *Electronic Government - 12th IFIP WG 8.5 International Conference, EGOV 2013, Koblenz, Germany, September 16-19, 2013. Proceedings*, volume 8074 of *Lecture Notes in Computer Science*, pages 99–110. Springer, 2013. doi: 10.1007/978-3-642-40358-3_9. URL https://doi.org/10.1007/978-3-642-40358-3_9.

Evangelos Kalampokis, Bill Roberts, Areti Karamanou, Efthimios Tambouris, and Konstantinos A. Tarabanis. Challenges on developing tools for exploiting linked open data cubes. In Sarven Capadisli, Franck Cotton, Armin Haller, Evangelos Kalampokis, Monica Scannapieco, and Raphaël Troncy, editors, *Proceedings of the 3rd International Workshop on Semantic Statistics co-located with 14th International Semantic Web Conference, SemStats@ISWC 2015, Bethlehem, Pennsylvania, USA, October 11th, 2015*, volume 1551 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015. URL http://ceur-ws.org/Vol-1551/article-07.pdf.

Ralph Kimball. *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. John Wiley, 1996. ISBN 0-471-15337-0.

Jakub Klímek, Jan Kucera, Martin Necaský, and Dusan Chlapek. Publication and usage of official czech pension statistics linked open data. *J. Web Semant.*, 48:1–21, 2018. doi: 10.1016/j.websem.2017.09.002. URL https://doi.org/10.1016/j.websem.2017.09.002.

Dimitris Kontokostas, Patrick Westphal, Sören Auer, Sebastian Hellmann, Jens Lehmann, and Roland Cornelissen. Databugger: a test-

driven framework for debugging the web of data. In Chin-Wan Chung, Andrei Z. Broder, Kyuseok Shim, and Torsten Suel, editors, *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 115–118. ACM, 2014. doi: 10.1145/2567948.2577017. URL https://doi.org/10.1145/2567948.2577017.

Yongju Lee. A life-cycle workflow architecture for linked data. In *Proceedings of the 2017 International Conference on Machine Learning and Soft Computing, ICMLSC 2017, Ho Chi Minh City, Vietnam, January 13-16, 2017*, pages 117–121, 2017. doi: 10.1145/3036290.3036302. URL https://doi.org/10.1145/3036290.3036302.

Georg J. P. Link, Kevin Lumbard, Kieran Conboy, Michael Feldman, Joseph Feller, Jordana George, Matt Germonprez, Sean P. Goggins, Debora Jeske, Gaye Kiely, Kristen Schuster, and Matt Willis. Contemporary issues of open data in information systems research: Considerations and recommendations. *CAIS*, 41:25, 2017. URL http://aisel.aisnet.org/cais/vol41/iss1/25.

Martin Lnenicka and Jitka Komarkova. Developing a government enterprise architecture framework to support the requirements of big and open linked data with the use of cloud computing. *Int J. Information Management*, 46:124–141, 2019. doi: 10.1016/j.ijinfomgt.2018.12.003. URL https://doi.org/10.1016/j.ijinfomgt.2018.12.003.

Alejandro Maté, Juan Trujillo, and John Mylopoulos. Key performance indicator elicitation and selection through conceptual modelling. In *Conceptual Modeling - 35th International Conference, ER 2016, Gifu, Japan, November 14-17, 2016, Proceedings*, pages 73–80, 2016. doi: 10.1007/978-3-319-46397-1_6. URL https://doi.org/10.1007/978-3-319-46397-1_6.

Junichi Matsuda, Akie Mizutani, Yu Asano, Dan Yamamoto, Hideaki Takeda, Ikki Ohmukai, Fumihiro Kato, Seiji Koide, Hiromu Harada, and Shoki Nishimura. Publication of statistical linked open data in japan. In *Semantic Technology - 8th Joint International Conference, JIST 2018, Awaji, Japan, November 26-28, 2018, Proceedings*, pages 307–319, 2018. doi: 10.1007/978-3-030-04284-4_21. URL `https://doi.org/10.1007/978-3-030-04284-4_21`.

Brian McBride. The resource description framework (rdf) and its vocabulary description language rdfs. In *Handbook on ontologies*, pages 51–65. Springer, 2004.

Massimo Mecella, Monica Scannapieco, Antonino Virgillito, Roberto Baldoni, Tiziana Catarci, and Carlo Batini. Managing data quality in cooperative information systems. In *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002 Irvine, California, USA, October 30 - November 1, 2002, Proceedings*, pages 486–502, 2002. doi: 10.1007/3-540-36124-3_28. URL `https://doi.org/10.1007/3-540-36124-3_28`.

Vuk Mijović, Valentina Janev, Dejan Paunović, and Sanja Vraneš. Exploratory spatio-temporal analysis of linked statistical data. *Journal of Web Semantics*, 41:1 – 8, 2016. ISSN 1570-8268. doi: https://doi.org/10.1016/j.websem.2016.10.002. URL `http://www.sciencedirect.com/science/article/pii/S1570826816300488`.

Mark A. Musen. The protégé project: a look back and a look forward. *AI Matters*, 1(4):4–12, 2015. doi: 10.1145/2757001.2757003. URL `https://doi.org/10.1145/2757001.2757003`.

Alberto Nogales, Miguel-Ángel Sicilia, Salvador Sánchez Alonso, and Elena García Barriocanal. Linking from schema.org microdata to

the web of linked data: An empirical assessment. *Computer Standards & Interfaces*, 45:90–99, 2016. doi: 10.1016/j.csi.2015.12.003. URL https://doi.org/10.1016/j.csi.2015.12.003.

Natalya F Noy, Deborah L McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

Adegboyega K. Ojo, Edward Curry, and Fatemeh Ahmadi Zeleti. A tale of open data innovations in five smart cities. In Tung X. Bui and Ralph H. Sprague Jr., editors, *48th Hawaii International Conference on System Sciences, HICSS 2015, Kauai, Hawaii, USA, January 5-8, 2015*, pages 2326–2335. IEEE Computer Society, 2015. doi: 10.1109/HICSS.2015.280. URL https://doi.org/10.1109/HICSS.2015.280.

Open Data Barometer. Global report. https://opendatabarometer.org/4thedition/report/, 2018. [Online; accessed 5-April-2018].

Open data charter. Open Up Field Guides. Methodology. https://drive.google.com/file/d/1itEjUOzSdn35K0o7VLoxrYzKHEwASYKV/view, 2018. [Online; accessed 24-February-2020].

Open Data Charter (ODC). International Open Data Charter. https://opendatacharter.net/wp-content/uploads/2015/10/opendatacharter-charter_F.pdf, 2015. [Online; accessed April-2019].

Open Data Charter (ODC). Open Up Field Guides. Methodology. https://drive.google.com/file/d/1itEjUOzSdn35K0o7VLoxrYzKHEwASYKV/view, 2018. [Online; April-2019].

Open Knowledge International. Open Data Handbook: File formats. http://opendatahandbook.org/guide/en/appendices/file-formats/, 2012. [Online; accessed 7-April-2018].

Zhengyu Pan, Tao Zhu, Hong Liu, and Huansheng Ning. A survey of RDF management technologies and benchmark datasets. *J. Ambient Intelligence and Humanized Computing*, 9(5):1693–1704, 2018. doi: 10.1007/s12652-018-0876-2. URL https://doi.org/10.1007/s12652-018-0876-2.

Karl Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A*, (187):253–318, 1896.

Leo Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Commun. ACM*, 45(4):211–218, 2002. doi: 10.1145/505248.5060010. URL http://doi.acm.org/10.1145/505248.5060010.

Alessandro Piscopo. Wikidata:Requests for comment/Data quality framework for Wikidata. https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Data_quality_framework_for_Wikidata, 2016. [Online; accessed 11-February-2018].

European Data Portal. Re-using Open Data. https://www.europeandataportal.eu/sites/default/files/re-using_open_data.pdf, 2017. [Online; accessed 9-April-2018].

Eric Prud, Andy Seaborne, et al. Sparql query language for rdf. *W3C recommendation*, 2006.

Jose Pérez, Francisco Ruiz, and Mario Piattini. Model driven engineering aplicado a business process management. 01 2007.

Monika Rani, Sanchit Alekh, Aditya Bhardwaj, Abhinav Gupta, and O. P. Vyas. Ontology-based classification and analysis of non-

emergency smart-city events. *CoRR*, abs/1708.00856, 2017. URL http://arxiv.org/abs/1708.00856.

RDF Working Group. Resource Description Framework (RDF). http://www.w3.org/RDF, 2014. [Online; accessed 15-November-2018].

Thomas Rebele, Fabian M. Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In Paul T. Groth, Elena Simperl, Alasdair J. G. Gray, Marta Sabou, Markus Krötzsch, Freddy Lécué, Fabian Flöck, and Yolanda Gil, editors, *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*, volume 9982 of *Lecture Notes in Computer Science*, pages 177–185, 2016. doi: 10.1007/978-3-319-46547-0_19. URL https://doi.org/10.1007/978-3-319-46547-0_19.

Mercedes Reguant Alvarez and Mercedes Fonseca. El método delphi-. *Reire*, 9:87–102, 01 2016. doi: 10.1344/reire2016.9.1916.

Xiangnan Ren, Olivier Curé, Li Ke, Jérémy Lhez, Badre Belabbess, Tendry Randriamalala, Yufan Zheng, and Gabriel Képéklian. Strider: An adaptive, inference-enabled distributed RDF stream processing engine. *PVLDB*, 10(12):1905–1908, 2017. doi: 10.14778/3137765.3137805. URL http://www.vldb.org/pvldb/vol10/p1905-ren.pdf.

Gustavo Candela Romero, Maria Pilar Escobar Esteban, Rafael C. Carrasco, and Manuel Marco Such. Migration of a library catalogue into RDA linked open data. *Semantic Web*, 9(4):481–491, 2018. doi: 10.3233/SW-170274. URL https://doi.org/10.3233/SW-170274.

Erna Ruijer, Stephan Grimmelikhuijsen, Michael J. Hogan, Sem Enzerink, Adegboyega Ojo, and Albert Meijer. Connecting societal

issues, users and data. scenario-based design of open data platforms. *Government Information Quarterly*, 34(3):470–480, 2017. doi: 10.1016/j.giq.2017.06.003. URL https://doi.org/10.1016/j.giq.2017.06.003.

Maheyand M.and Al-Abdullaand A.and Amesand S.and Brayand P.and Candelaand G.and Chambersand S.and Dervenand C.and Dobreva-McPhersonand M.and Gasserand K.and Karnerand S.and Kokegeiand K.and Laursenand D.and Potterand A.and Straubeand A.and Wagnerand S-C., Wilmsand L.and with forewords by: Al-Emadiand T. A.and Broady-Prestonand J.and Landryand P., and Papaioannouand G. *Open a Glam Lab*. Digital Cultural Heritage Innovation Labs, Book Sprint, 2019.

Leo Sauermann, Richard Cyganiak, Danny Ayers, and Max Völkel. Cool URIs for the Semantic Web. W3C Interest Group Note 03 December 2008. https://www.w3.org/TR/cooluris/, 2008. [Online; accessed 5-April-2018].

S. S. SHAPIRO and M. B. WILK. An analysis of variance test for normality (complete samples)†. *Biometrika*, 52(3-4):591–611, 12 1965. ISSN 0006-3444. doi: 10.1093/biomet/52.3-4.591. URL https://doi.org/10.1093/biomet/52.3-4.591.

Sarah L. Shreeves, Ellen Knutson, Besiki Stvilia, Carole L. Palmer, Michael B. Twidale, and Timothy W. Cole. Is quality metadata shareable metadata? the implications of local metadata practices for federated collections. 2005.

Aikaterini-Maria Sourouni, Giorgos Kourlimpinis, Spiros Mouzakitis, and Dimitris Askounis. Towards the government transformation: An ontology-based government knowledge repository. *Computer Standards & Interfaces*, 32(1-2):44–53, 2010. doi: 10.1016/j.csi.2009.06.002. URL https://doi.org/10.1016/j.csi.2009.06.002.

Carl Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, (15):72–101, 1904.

Steffen Staab and Rudi Studer. *Handbook on ontologies*. Springer Science & Business Media, 2013.

Thomas Pellissier Tanon, Denny Vrandecic, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. From freebase to wikidata: The great migration. In Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao, editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 1419–1428. ACM, 2016. doi: 10.1145/2872427.2874809. URL `https://doi.org/10.1145/2872427.2874809`.

The World Bank. Starting an Open Data Initiative. `http://opendatatoolkit.worldbank.org/en/starting.html`, 2013. [Online; accessed 5-April-2018].

Gaëlle Thivet and Sara Fernandez. *Water Demand Management: The Mediterranean Experience. Technical focus paper.* Global Water Partnership (GWP), 2012. ISBN 978-91-85321-88-9.

Tim Berners-Lee. Linked Data. `https://www.w3.org/DesignIssues/LinkedData.html`, 2006. [Online; accessed 5-April-2018].

Jovan Varga, Alejandro A. Vaisman, Oscar Romero, Lorena Etcheverry, Torben Bach Pedersen, and Christian Thomsen. Dimensional enrichment of statistical linked open data. *Journal of Web Semantics*, 40:22 – 51, 2016. ISSN 1570-8268. doi: https://doi.org/10.1016/j.websem.2016.07.003. URL `http://www.sciencedirect.com/science/article/pii/S1570826816300348`.

Daniel Vila-Suero, Boris Villazón-Terrazas, and Asunción Gómez-Pérez. datos.bne.es: A library linked dataset. *Semantic Web*, 4 (3):307–313, 2013. doi: 10.3233/SW-120094. URL https://doi.org/10.3233/SW-120094.

Boris Villazón-Terrazas, Luis. M. Vilches-Blázquez, Oscar Corcho, and Asunción Gómez-Pérez. *Methodological Guidelines for Publishing Government Linked Data*, chapter–, pages 27–49. Springer New York, New York, NY, 2011. ISBN 978-1-4614-1767-5. doi: 10.1007/978-1-4614-1767-5_2. URL https://doi.org/10.1007/978-1-4614-1767-5_2.

W3C. Inference. https://www.w3.org/standards/semanticweb/inference. [Online; accessed 20-November-2018].

W3C. Notation3 (n3): A readable rdf syntax. "https://www.w3.org/TeamSubmission/n3/", 2011. [Online; accessed 13-November-2018].

Andra Waagmeester, Egon L. Willighagen, Núria Queralt-Rosinach, Elvira Mitraka, Sebastian Burgstaller-Muehlbacher, Tim E. Putman, Julia Turner, Lynn M. Schriml, Paul Pavlidis, Andrew I. Su, and Benjamin M. Good. Linking wikidata to the rest of the semantic web. In *Proceedings of the 9th International Conference Semantic Web Applications and Tools for Life Sciences, Amsterdam, The Netherlands, December 5-8, 2016.*, 2016. URL http://ceur-ws.org/Vol-1795/paper46.pdf.

Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *J. of Management Information Systems*, 12(4):5–33, 1996. URL http://www.jmis-web.org/articles/1002.

Wikidata. SPARQL federation input/Archive. https://www.wikidata.org/wiki/Wikidata:SPARQL_federation_input/Archive, 2017. [Online; accessed 10-July-2018].

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018–, 2016. doi: 10.1038/sdata.2016.18. URL https://doi.org/10.1038/sdata.2016.18.

Working Group Note. Best Practices for Publishing Linked Data. https://www.w3.org/TR/ld-bp/, 2014. [Online; accessed 20-May-2018].

World Wide Web Consortium (W3C). OWL Web Ontology Language Overview. http://www.w3.org/TR/owl-features/, 2004. [Online; accessed 7-April-2018].

World Wide Web Consortium (W3C). Semantic Integration & Interoperability Using RDF and OWL. https://www.w3.org/

2001/sw/BestPractices/OEP/SemInt/, 2005. [Online; accessed 04-September-2019].

World Wide Web Consortium (W3C). Describing linked datasets with the void vocabulary. https://www.w3.org/TR/void/, 2011. [Online; accessed 19-June-2018].

World Wide Web Consortium (W3C). PROV-O: The PROV Ontology. https://www.w3.org/TR/prov-o/, 2013. [Online; accessed 1-August-2018].

World Wide Web Consortium (W3C). RDF Schema 1.1. https://www.w3.org/TR/rdf-schema/, 2014. [Online; accessed 7-April-2018].

World Wide Web Consortium (W3C). Data on the Web Best Practices. https://www.w3.org/TR/dwbp/, 2017. [Online; accessed 19-June-2018].

World Wide Web Consortium (W3C). Comparison of rdfjs libraries. https://www.w3.org/community/rdfjs/wiki/Comparison_of_RDFJS_libraries, 2018. [Online; accessed 29-November-2018].

Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016. doi: 10.3233/SW-150175. URL https://doi.org/10.3233/SW-150175.

A. Ontología para la red de abastecimiento de agua

A continuación se describe con más detalle la ontología utilizada para la descripción del dominio de la red de abastecimiento de agua potable. Esta ontología ha sido creada por María del Mar Roldán-García y José García-Nieto, Departamento de Lenguajes y Ciencias de la Computación, Ada Byron Research Building, Universidad de Málaga, España, en colaboración con nuestro grupo.

Uno de los principales objetivos en el paso de generación del RDF es capturar, consolidar e integrar datos de diferentes fuentes relacionadas con el caso de uso real, los datos del agua. Por esta razón, en este primer experimento se optó por diseñar un enfoque semántico para el intercambio de datos y la reconciliación, mediante el cual se utiliza un modelo ontológico para agrupar el conocimiento común del dominio en el que se sitúa la experimentación. Para ello se ha desarrollado una ontología OWL para describir los indicadores utilizados por la empresa de suministro de agua. Los términos más relevantes de la ontología se obtuvieron directamente de los tomadores de decisiones de la compañía. La ontología está compuesta de 31 clases, 23 propiedades de objeto, 2 propiedades de datos, 156 axiomas lógicos y 12 individuales.

A continuación se describen las clases principales *Indicator*, *zone* y las propiedades más destacables.

`Indicator`, son los indicadores suministrados por la empresa del agua en los datos originales y en este caso representan medidas (*measures* miden actividades comerciales). Cada indicador tiene un valor en una unidad de medida y en un año en concreto (la tabla A.1 muestra

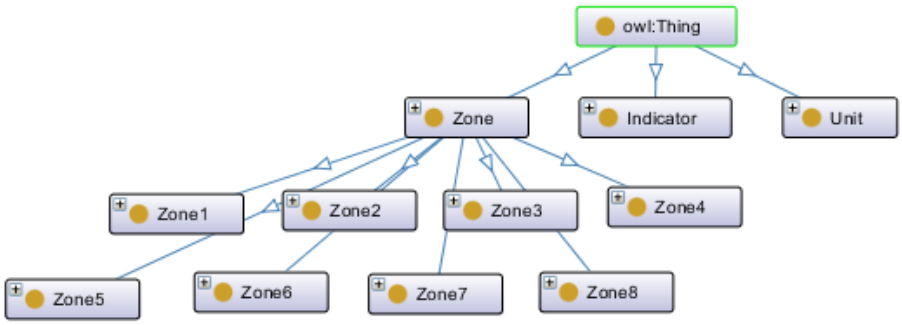


Figura A.1: Clases principales de la ontología. Fuente *An Ontology-Based Framework for Publishing and Exploiting Linked Open Data: A Use Case on Water Resources Management.*

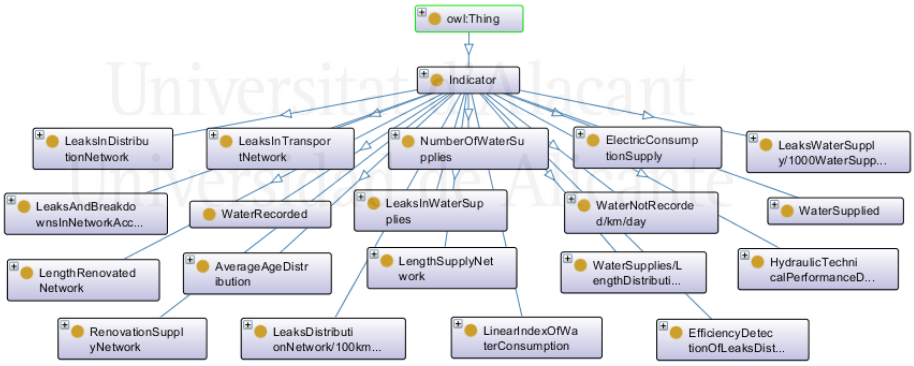


Figura A.2: Subclases de `Indicator`. Fuente *An Ontology-Based Framework for Publishing and Exploiting Linked Open Data: A Use Case on Water Resources Management.*

Tabla A.1: *Indicator* objeto y propiedades. Fuente *An Ontology-Based Framework for Publishing and Exploiting Linked Open Data: A Use Case on Water Resources Management.*

Propiedades de objeto	Descripción lógica
unit	$\exists$ unit.Thing $\sqsubseteq$ Indicator
	$\top \sqsubseteq \forall$ unit.Unit
Propiedades de datos	**Descripción lógica**
value	$\exists$ value.Datatype Literal $\sqsubseteq$ Indicator
	$\top \sqsubseteq \forall$ value.Datatype double
inYear	$\exists$ inYear.Datatype Literal $\sqsubseteq$ Indicator
	$\top \sqsubseteq \forall$ inYear.Datatype integer

las propiedades de objeto y datos). Las subclases del indicador son: `NumberOfWaterSupplies`, `LeaksTansportNetwork/100kmNetwork`, `AverageAgeDistribution`, `LengthRenovatedNetwork` , etc. (ver la figura A.2)

La clase `Unit` contiene las diferentes unidades de medida, por ejemplo: `Km`, $\mathtt{m}^3$ y `number/1000WaterSupplies`. La clase `Year` contiene los diferentes años.

`Zone` modela las zonas donde los indicadores se miden y cogen valor. Cada zona se divide en subzonas modeladas como individuos de la clase `Zone`. La ontología tiene 8 zonas (`Zone1-Zone8`) que se dividen en subzonas, por ejemplo `Zone2` tiene 14 subzonas (`zone2.1-zone2.14`). Cada indicador se relaciona con las subzonas donde coge el valor (es medido) y se describe a través de la propiedad de objeto `hasIndicator` y sus subclases como `hasNumberOfWaterSupplies`, `hasHydraulicTechnicalPerformanceDistribution`, `hasLeaksTansportNetwork/100kmNetwork`. Por lo tanto, una zona concreta está relacionada con una subclase de `Indicator` por medio de la propiedad `hasIndicator` y se especifican su valor, unidad de medida y año.

La figura A.3 muestra una representación gráfica de estas relaciones

Tabla A.2: Modelado *Hydraulic technical performance distribution.*

ObjectProperty	Description Logic
hasHydraulicTechnicalPerformanceDistribution	$\sqsubseteq$ hasIndicator $\exists$ hasHydraulicTechnicalPerformanceDistribution.Thing $\sqsubseteq$ Zone $\top \sqsubseteq \forall$ hasHydraulicTechnicalPerformanceDistribution .HydraulicTechnicalPerformanceDistribution
classes	Description Logic
HydraulicTechnicalPerformanceDistribution	$\sqsubseteq$ Indicator

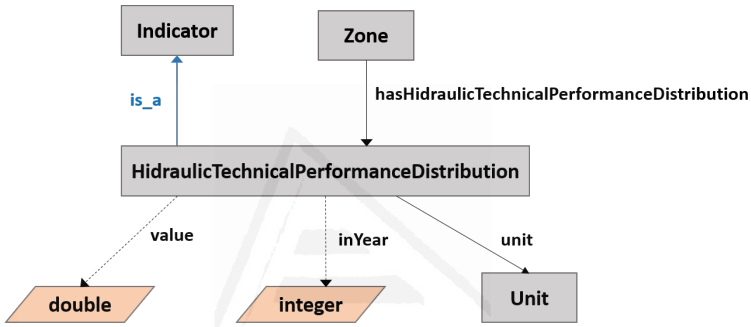


Figura A.3: Relaciones entre *Zones* y *Indicators*, ejemplo del modelado *HydraulicTechnicalPerformanceDistribution* modeling. Fuente *An Ontology-Based Framework for Publishing and Exploiting Linked Open Data: A Use Case on Water Resources Management.*

para el indicador "Distribución del rendimiento técnico hidráulico", mientras que la tabla A.2 presenta los axiomas lógicos utilizados en la ontología. La figura A.4 muestra un ejemplo de cómo la medida del indicador "Distribución del rendimiento técnico hidráulico" para la subzona 2.4 en 2009 se modela siguiendo los axiomas de ontología. El resto de indicadores se modelan de manera similar.

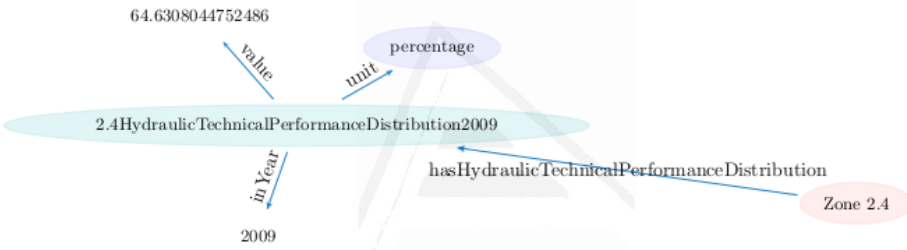


Figura A.4: Representación del grafo según la ontología descrita. Fuente *An Ontology-Based Framework for Publishing and Exploiting Linked Open Data: A Use Case on Water Resources Management.*

B. Ejemplo de fichero VoID

Fichero VoID para el repositorio *Linked Data version of the critical cleaning spots in Barcelone* creado en la segunda iteración.

```
@prefix void: <http://rdfs.org/ns/void#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.

<> a void:DatasetDescription;
    dcterms:title "A VoID Description of the RDF Data Cube
        Dataset";
    dcterms:creator <https://smartdataua.github.io/
        rdfdatacube/>;
    foaf:homepage <https://smartdataua.github.io/
        rdfdatacube/>;
    .

:dataset rdf:type void:Dataset ;
 foaf:homepage <https://smartdataua.github.io/rdfdatacube/>
     ;
 dcterms:title "A VoID Description of the RDF Data Cube
     Dataset" ;
 dcterms:description "Linked Data version of the critical
     cleaning spots in Barcelone" ;
 dcterms:publisher :University_Alicante;
 dcterms:creator :gcandela;
 dcterms:creator :pescobar;
 dcterms:modified "2018-06-26"^^xsd:date;
 void:vocabulary <http://purl.org/dc/elements/1.1/> ;
 void:vocabulary <http://www.w3.org/2004/02/skos/core#> ;
 void:vocabulary <http://www.w3.org/2000/01/rdf-schema#> ;
 void:vocabulary <http://purl.org/linked-data/cube#> ;
```

```
void:exampleResource <http://example.com/201705/obs1> ;
void:uriSpace "http://example.com/";
dcterms:subject <https://www.wikidata.org/wiki/Q1492>;
dcterms:subject <https://www.wikidata.org/wiki/Q58734>;
void:feature <http://www.w3.org/ns/formats/RDF_XML>;
.

:University_Alicante a foaf:Organization;
    rdfs:label "Universidad de Alicante";
    foaf:homepage <http://www.ua.es>;
.
:gcandela a foaf:Person;
    rdfs:label "Gustavo Candela";
    foaf:mbox <mailto:gcandela@dlsi.ua.es>;
    .
:pescobar a foaf:Person;
    rdfs:label "Pilar Escobar";
    foaf:mbox <mailto:pescobar@dlsi.ua.es>;
    .
```

C. Cálculo del coeficiente de competencia experta

El coeficiente de competencia experta K se calcula a partir de la autovaloración realizada por cada uno de los expertos [Cabrero Almenara and Barroso Osuna, 2013]. Para calcular este coeficiente se aplica la siguiente fórmula $K = \frac{1}{2}(Kc + Ka)$, donde Kc es el coeficiente de conocimiento, es decir, la información que previamente tiene el experto sobre la cuestión que se le plantea y se calcula según la propia valoración del experto en una escala de 0 a 10. Para calcular el kc, al valor obtenido se multiplica por $0,1$. Y Ka es el coeficiente de argumentación, es decir, la fundamentación de los criterios de los expertos, se calcula en función de la valoración de un conjunto de fuentes de argumentación. La tabla C.1 muestra los valores otorgados usualmente a cada una de las fuentes de argumentación según su valor alto, medio o bajo. El anexo D se puede consultar el test enviado al grupo de expertos para su propia valoración.

La investigación va dirigida a nuevos modelos para la publicación y reutilización de los datos abiertos enlazados, teniendo en cuenta las diferentes fases por las que debe pasar el conjunto de datos para asegurar a los usuarios datos reutilizables y fiables.

La tabla C.2 muestra los valores obtenidos en la autoevaluación del grupo de expertos para cada una de las fuentes de argumentación listadas del 1 al 6 en la tabla C.1, el coeficiente de conocimiento kc, el coeficiente de argumentación ka y por último el coeficiente de competencia experta K. En función del valor obtenido en k y calculado según la fórmula $K = \frac{1}{2}(Kc + Ka)$ se clasifican los expertos en: alta, media

Tabla C.1: Valoración de las fuentes de argumentación. Fuente: [Cabrero Almenara and Barroso Osuna, 2013]

Fuente de argumentación	Grado de influencia de cada una de las fuentes en sus criterios		
	Alto	Medio	Bajo
1.Análisis teóricos realizados	0,3	0,2	0,1
2.Experiencia obtenida	0,5	0,4	0,2
3.Estudio de trabajos de autores españoles	0,05	0,05	0,05
4.Estudio de trabajos de autores internacionales	0,05	0,05	0,05
5.Conocimiento sobre el estado del problema en el extranjero	0,05	0,05	0,05
6.Intuición personal del experto	0,05	0,05	0,05

y baja influencia. Según la literatura en el cálculo de k, los expertos cuyo valor sea inferior a $0,8$ no deberían formar parte del grupo de expertos.

- $k > 0,8$ tiene una alta influencia en todas las fuentes.
- $0,7 \leq k \leq 0,8$ tiene una influencia media en todas las fuentes.
- $0,5 \leq k < 0,7$ tiene una influencia baja en todas las fuentes.

Catorce de los expertos obtuvieron un índice K mayor o igual a $0,8$ (seis de cuales obtuvieron el máximo valor, 1) solo uno de los expertos obtuvo un valor por debajo de $0,8$ en concreto, un valor de $0,75$. Con estos datos, los quince expertos fueron considerados adecuados para realizar la valoración del modelo. Diez de los expertos son doctores en informática ($66,6\,\%$) y catorce de ellos realiza tareas de docencia

Tabla C.2: Tabla resumen de los valores obtenidos por los expertos en las fuentes de argumentación 1-6, coeficiente de conocimiento kc, coeficiente de argumentación ka y el coeficiente de competencia experta K

Experto	1	2	3	4	5	6	kc	ka	K
1	0,3	0,5	0,05	0,05	0,05	0,05	1	1	1
2	0,2	0,4	0,05	0,05	0,05	0,05	0,8	0,8	0,8
3	0,3	0,5	0,05	0,05	0,05	0,05	1	1	1
4	0,2	0,4	0,05	0,05	0,05	0,05	0,8	0,8	0,8
5	0,3	0,5	0,05	0,05	0,05	0,05	1	1	1
6	0,2	0,4	0,05	0,05	0,05	0,05	0,8	0,8	0,8
7	0,3	0,4	0,05	0,05	0,05	0,05	0,7	0,9	0,8
8	0,3	0,5	0,05	0,05	0,05	0,05	1	1	1
9	0,2	0,5	0,05	0,05	0,05	0,05	0,7	0,9	0,8
10	0,3	0,5	0,05	0,05	0,05	0,05	0,9	1	1
11	0,2	0,4	0,05	0,05	0,05	0,05	0,7	0,8	0,75
12	0,3	0,5	0,05	0,05	0,05	0,05	1	1	1
13	0,2	0,4	0,05	0,05	0,05	0,05	0,8	0,8	0,8
14	0,3	0,5	0,05	0,05	0,05	0,05	0,8	1	0,9
15	0,3	0,5	0,05	0,05	0,05	0,05	0,9	1	0,95

e investigación. La muestra recoge expertos de edades comprendidas entre 27 y 63 años y que 13,3 % son mujeres.

D. Cuestionario de autovaloración para el grupo de expertos

A continuación se muestra el test realizado al grupo de expertos para su autovaloración. Los resultados obtenidos en este test se han procesado para calcular el coeficiente de competencia experta.

> Gracias por colaborar en este trabajo de investigación, por favor, se le solicita que realice el siguiente test de valoración personal sobre su conocimiento del tema tratado en esta investigación.

1. Entidad en la que trabaja
2. Grado científico
   - Doctor
   - Máster
   - Ingeniería
   - Otros
3. Categoría docente
   - Catedrático
   - Profesor titular
   - Contratado doctor
   - Ayudante doctor
   - Profesor asociado
   - No realizo tareas de personal docente e investigador
   - Soy investigador pero no trabajo como docente

4. Edad
5. Género
    - Femenino
    - Masculino
    - Prefiero no contestar

Por favor, indique su grado de conocimientos acerca del tema de investigación de este experimento "Un enfoque multidimensional basado en RDF para la publicación de Linked Open Data", valorando en una escala ascendente del 1 al 10.

Fuentes de argumentación. Por favor, indique el grado de influencia que cada una de las fuentes que a continuación se enumeran ha tenido en su conocimiento y los criterios que posee sobre el tema de investigación

1. Análisis teóricos realizados
    - Alto
    - Medio
    - Bajo
2. Experiencia obtenida
    - Alto
    - Medio
    - Bajo
3. Estudio de trabajos de autores españoles
    - Alto
    - Medio
    - Bajo
4. Estudio de trabajos de autores internacionales

- Alto
- Medio
- Bajo

5. Conocimiento sobre el estado del problema en el extranjero
   - Alto
   - Medio
   - Bajo
6. Intuición personal del experto
   - Alto
   - Medio
   - Bajo

E. Valoración mediante criterio de expertos. Iteración 1

Existen dos aspectos importantes a tratar cuando se trabaja con el método Delphi. El primero es qué criterios marcarán la selección del grupo de expertos y el segundo qué número de expertos es el adecuado para llevar a cabo la validación. Como ya se ha comentado en el capítulo 2 se ha seleccionado directamente un grupo de expertos en función de su vinculación y experiencia profesional con la temática a evaluar. En el anexo C se ha calculado el coeficiente de competencia experta que demuestra que el grupo de expertos es adecuado para realizar la validación del modelo propuesto. Se han seleccionado 15 expertos para realizar esta valoración.

Para poder valorar la opinión de los expertos sobre el modelo se ha utilizado la escala de Likert[1]. Esta técnica permite medir el grado de aceptación o no del modelo por parte de los expertos. Para ello se ha diseñado una encuesta que permite identificar el grado de acuerdo o desacuerdo en cada uno de los indicadores o cuestiones planteadas, normalmente se utilizan 5 niveles para realizar esta medición:

1 Completamente en desacuerdo

2 En desacuerdo

3 Ni de acuerdo ni en desacuerdo

4 De acuerdo

[1] `https://en.wikipedia.org/wiki/Likert_scale`

5 Muy de acuerdo

Con ello se pretende obtener la opinión de los expertos como apoyo para una mejora en el modelo. Como se ha descrito en el capítulo 2 en que se ha descrito la metodología de la investigación y el planteamiento de dos iteraciones para refinar el modelo propuesto.

En el anexo F se puede consultar la encuesta realizada al grupo de expertos para valorar la primera propuesta del modelo (iteración 1). Se les ha suministrado un resumen del modelo y descripción de los pasos que lo componen para que puedan evaluarlo así como los principios fundamentales, además de realizar entrevistas para la resolución de dudas.

La tabla E.1 muestra los valores obtenidos en las encuestas realizadas a los expertos. Para cada indicador se ha calculado el número de respuestas y su porcentaje según los diferentes niveles en la escala de Likert, estos niveles vienen identificados por:

CD Completamente en desacuerdo

ED En desacuerdo

N-Si-No- Ni de acuerdo ni en desacuerdo

DA De acuerdo

MA Muy de acuerdo

Los indicadores valorados por los expertos y cuyos resultados se muestran en la tabla E.1, son los siguientes:

1. Interoperabilidad
2. Reutilización
3. Calidad

4. Enriquecimiento
5. Valore la fase de especificación de las fuentes de datos en la que se realiza mapeo y preprocesamiento de fuentes
6. ¿Cómo valora el modelado y generación del RDF?
7. Valore la publicación del conjunto de datos con acceso público
8. ¿Cómo valora la explotación del conjunto de datos?

La última columna de la tabla E.1 muestra el valor del IP obtenido (índice porcentual) y representa el grado de acuerdo de los expertos para ese determinado indicador. El índice porcentual se ha calculado según la fórmula E.1.

$$IP = \frac{5(\%MA)+4(\%DA)+3(\%N-Si-No)+2(\%ED)+1(\%CD)}{5} \qquad \text{(E.1)}$$

Tabla E.1: Tabla resumen de los valores recogidos de la encuesta realizada al grupo de expertos para la primera iteración del modelo. Cada fila representa los valores obtenidos en cada una de las cuestiones planteadas en la encuesta, según la escala de Likert.

	CD		ED		N-Si-No		DA		MA		
Ítems	Núm.	%	Núm.	%	Núm.	%	Núm.	%	Núm.	%	IP
1	0	0,00	0	0,00	0	0,00	5	33,33	10	66,67	93,33
2	0	0,00	0	0,00	1	6,67	8	53,33	6	40,00	86,67
3	5	33,33	5	33,33	5	33,33	0	0,00	0	0,00	40,00
4	0	0,00	0	0,00	0	0,00	4	26,67	11	73,33	94,67
5	0	0,00	0	0,00	0	0,00	5	33,33	10	66,67	93,33
6	0	0,00	0	0,00	0	0,00	3	20,00	12	80,00	96,00
7	0	0,00	0	0,00	0	0,00	6	40,00	9	60,00	92,00
8	4	26,67	1	6,67	0	0,00	3	20,00	7	46,67	70,67

F. Formulario para la valoración de la iteración 1

A continuación se muestra la encuesta realizada al grupo de expertos para valorar la primera propuesta del modelo (iteración 1). Se les ha suministrado un resumen del modelo y descripción de los pasos que lo componen para que puedan evaluarlo así como los principios fundamentales. También se han realizado entrevistas para la resolución de dudas sobre el modelo.

> Esta encuesta pretende recoger su valoración de experto en los aspectos fundamentales propuestos en esta investigación "Un enfoque multidimensional basado en RDF para la publicación de Linked Open Data". Muchas gracias por su colaboración.

1. Entidad en la que trabaja
2. Grado científico
   - Doctor
   - Máster
   - Ingeniería
   - Otros
3. Categoría docente
   - Catedrático
   - Profesor titular
   - Contratado doctor

- Ayudante doctor
- Profesor asociado
- No realizo tareas de personal docente e investigador
- Soy investigador pero no trabajo como docente

4. Edad
5. Género
    - Femenino
    - Masculino
    - Prefiero no contestar

Por favor, valore los principios fundamentales asociados al modelo. Valorando de 1 a 5 si cree que el modelo refleja estos principios siendo 1 el valor más bajo en el que está completamente en desacuerdo (el modelo no lo contempla) y 5 en el que está muy de acuerdo (está totalmente integrado).

1. Interoperabilidad
2. Reutilización
3. Calidad
4. Enriquecimiento
5. Valore la fase de especificación de las fuentes de datos en la que se realiza mapeo y preprocesamiento de fuentes
6. ¿Cómo valora el modelado y generación del RDF?
7. Valore la publicación del conjunto de datos con acceso público.
8. ¿Cómo valora la explotación del conjunto de datos?

A continuación puede indicar algún comentario o sugerencia que considere necesaria sobre el modelo propuesto.

G. Valoración mediante criterio de expertos. Iteración 2

Después de refinar el modelo y aplicarlo a un caso real se volvió a realizar la valoración por parte de los expertos. Al igual que en la primera iteración se han procesados los resultados obtenidos de dicha valoración siguiendo el mismo método descrito en el anexo E.

La tabla G.1 muestra los resultados obtenidos en cada indicador valorado por el experto. Para cada indicador se ha calculado el número de respuestas, su porcentaje según los diferentes niveles en la escala de Likert y el índice porcentual que representa el grado de acuerdo de los expertos para ese determinado indicador (calculado según la fórmula E.1).

Los indicadores valorados por los expertos en la segunda iteración son los siguientes:

1. Interoperabilidad
2. Reutilización
3. Calidad
4. Enriquecimiento
5. Valore la fase de especificación de las fuentes de datos en la que se realiza mapeo y preprocesamiento de fuentes
6. Valore el paso de modelado de datos

7. ¿Cómo valora el paso de generación del RDF?
8. ¿Cómo valoraría la publicación del conjunto de datos con acceso público?
9. ¿Se realiza la validación de calidad de los datos?
10. ¿Cómo valora la explotación del conjunto de datos?

La encuesta realizada al grupo de expertos para valorar el modelo en la iteración 2 se puede consultar en el anexo H. Al igual que en la valoración de la iteración 1, se les ha suministrado un resumen del modelo y descripción de los pasos que lo componen así como los principios fundamentales.

Tabla G.1: Tabla resumen de los valores recogidos de la encuesta realizada al grupo de expertos para la segunda iteración del modelo. Cada fila representa los valores obtenidos en cada una de las cuestiones planteadas en la encuesta, según la escala de Likert.

	CD		ED		N-Si-No		DA		MA		
Ítems	Núm.	%	Núm.	%	Núm.	%	Núm.	%	Núm.	%	IP
1	0	0,00	0	0,00	0	0,00	5	33,33	10	66,67	93,33
2	0	0,00	0	0,00	0	0,00	2	13,33	13	86,67	97,33
3	0	0,00	0	0,00	1	6,67	5	33,33	9	60,00	90,67
4	0	0,00	0	0,00	0	0,00	2	13,33	13	86,67	97,33
5	0	0,00	0	0,00	0	0,00	5	33,33	10	66,67	93,33
6	0	0,00	0	0,00	0	0,00	2	13,33	13	86,67	97,33
7	0	0,00	0	0,00	0	0,00	3	20,00	12	80,00	96,00
8	0	0,00	0	0,00	0	0,00	2	13,33	13	86,67	97,33
9	0	0,00	0	0,00	0	0,00	6	40,00	9	60,00	92,00
10	0	0,00	0	0,00	3	20,00	3	20,00	9	60,00	88,00

H. Formulario para la valoración de la iteración 2

Esta encuesta pretende recoger su valoración de experto en los aspectos fundamentales propuestos en esta investigación "'Un enfoque multidimensional basado en RDF para la publicación de Linked Open Data". Muchas gracias por su colaboración.

1. Entidad en la que trabaja
2. Grado científico
   - Doctor
   - Máster
   - Ingeniería
   - Otros
3. Categoría docente
   - Catedrático
   - Profesor titular
   - Contratado doctor
   - Ayudante doctor
   - Profesor asociado
   - No realizo tareas de personal docente e investigador
   - Soy investigador pero no trabajo como docente
4. Edad
5. Género

- Femenino
- Masculino
- Prefiero no contestar

Por favor, valore los principios fundamentales asociados al modelo. Valorando de 1 a 5 si cree que el modelo refleja estos principios siendo 1 el valor más bajo en el que está completamente en desacuerdo (el modelo no lo contempla) y 5 en el que está muy de acuerdo (está totalmente integrado).

1. Interoperabilidad
2. Reutilización
3. Calidad
4. Enriquecimiento
5. Valore la fase de especificación de las fuentes de datos en la que se realiza mapeo y preprocesamiento de fuentes
6. Valore el paso de modelado de datos
7. ¿Cómo valora el paso de generación del RDF?
8. ¿Cómo valoraría la publicación del conjunto de datos con acceso público?
9. ¿Se realiza la validación de calidad de los datos?
10. ¿Cómo valora la explotación del conjunto de datos?

A continuación puede indicar algún comentario o sugerencia que considere necesaria sobre el modelo propuesto.

I. Cuestionario para la valoración mediante criterio de usuarios

A continuación se presenta el cuestionario utilizado para la evaluación mediante el criterio de usuarios. La primera parte recoge información administrativa y tres preguntas de control sobre el problema que se estudia. El segundo bloque corresponde a un cuestionario de 9 preguntas sobre la información suministrada y cuatro cuestiones más para valorar la experiencia.

Con esta encuesta se pretende valorar en qué grado los datos suministrados a los usuarios son útiles para ellos y que les puede ayudar a extraer conclusiones.

Los datos puede consultarlos en las siguientes direcciones:

- https://docs.google.com/spreadsheets/d/1JmfSKKZTs-vTq8RYGx2vRTdxEDqNrOV_MC8u6yy-pPE/edit?usp=sharing
- https://docs.google.com/spreadsheets/d/1neA6Fs4iHCRI1RPxcaac3YRreLlDJm6U7n528mg6sKE/edit?usp=sharing

Gracias por colaborar en este trabajo de investigación, por favor, se le solicita que conteste a las siguientes preguntas.

1. Entidad en la que trabaja
2. Estudios realizados
3. Edad
4. Género
    a) Femenino

b) Masculino

c) Prefiero no contestar

5. En una escala del 1 al 5, cómo valoraría sus conocimientos sobre datos abiertos, siendo 1 el valor más bajo (no los conozco , completamente en desacuerdo) y 5 en el más alto (experto en datos abiertos, muy de acuerdo).

   a) Completamente en desacuerdo

   b) En desacuerdo

   c) Ni de acuerdo ni en desacuerdo

   d) De acuerdo

   e) Muy de acuerdo

6. En una escala del 1 al 5, cómo valoraría sus conocimientos sobre el lenguaje de interrogación de repositorios SPARQL, siendo 1 el valor más bajo (no lo conozco , completamente en desacuerdo) y 5 en el más alto (experto en SPARQL, muy de acuerdo).

   a) Completamente en desacuerdo

   b) En desacuerdo

   c) Ni de acuerdo ni en desacuerdo

   d) De acuerdo

   e) Muy de acuerdo

7. En una escala del 1 al 5, cómo valoraría su experiencia con hojas de cálculo, por ejemplo en Excel, siendo 1 el valor más bajo (no he trabajado con hojas de cálculo ni las conozco, completamente en desacuerdo) y 5 en el más alto (experto en hojas de cálculo, muy de acuerdo).

a) Completamente en desacuerdo

b) En desacuerdo

c) Ni de acuerdo ni en desacuerdo

d) De acuerdo

e) Muy de acuerdo

A continuación se le pide que conteste a 9 cuestiones sobre la información que se le ha suministrado.

1. Según los datos suministrados, ¿podría indicar qué barrio de Barcelona es el que más puntos críticos de limpieza tiene?

   a) El Raval

   b) Les Corts

   c) No se puede extraer esa información de los datos suministrados

2. Según los datos suministrados, ¿podría indicar qué barrio de Barcelona es el que menos puntos críticos de limpieza tiene?

   a) Sant Antoni

   b) La Bordeta

   c) No se puede extraer esa información de los datos suministrados

3. Según los datos suministrados, ¿podría indicar cuántos puntos críticos tiene El Poble-sec?

   a) Aproximadamente 40

   b) Aproximadamente 10

c) No se puede extraer esa información de los datos suministrados.

4. Y según los datos ¿podría indicar cuánta población tiene el barrio El Poble-sec?

   a) Más de 45.000 habitantes
   b) Aproximadamente 40.000 habitantes
   c) No se puede extraer esa información de los datos suministrados

5. Según los datos suministrados, ¿se podría decir que el barrio que más población tiene es el que más puntos críticos de limpieza tiene?

   a) Si
   b) No
   c) No se puede extraer esa información

6. Según los datos suministrados, ¿se podría decir que el barrio que más población tiene es el que más renta per cápita tiene?

   a) Si
   b) No
   c) No se puede extraer esa información

7. Según los datos suministrados, ¿se podría decir que el barrio que más población tiene es el que menos renta per cápita tiene?

   a) Si
   b) No
   c) No se puede extraer esa información

8. Según los datos, el barrio de mayor renta per cápita...

    a) Está entre los 5 que más población tiene
    b) Está entre los 5 que menos población tiene
    c) No se puede extraer esa información

9. Según los datos sumnisitrados ¿cree que existe una relación entre los puntos críticos y la cantidad de población de un barrio? Es decir que el barrio que más población tiene se producen más puntos críticos de limpieza

    a) Si
    b) No
    c) No se puede extraer esa información

En base a cómo se le han presentado la información sobre los puntos críticos de limpieza en la ciudad de Barcelona, población y renta per cápita, valore en una escala del 1 al 5, siendo 1 la valoración más baja (completamente en desacuerdo) y 5 la máxima (muy de acuerdo), su experiencia en función de si le ha resultado sencillo encontrar la información que se le pedía en cada pregunta.

1. Completamente en desacuerdo
2. En desacuerdo
3. Ni de acuerdo ni en desacuerdo
4. De acuerdo
5. Muy de acuerdo

En base a la información facilitada, valore en una escala del 1 al 5, siendo 1 la valoración más baja (completamente en desacuerdo) y 5 la máxima (muy de acuerdo), si le ha resultado útil para responder a las cuestiones planteadas.

1. Completamente en desacuerdo
2. En desacuerdo
3. Ni de acuerdo ni en desacuerdo
4. De acuerdo
5. Muy de acuerdo

Si necesitara reutilizar esta información qué grado de importancia le daría a la evaluación de la calidad de los datos que se publican (que provengan de fuentes fiables, datos completos, accesibles, etc.), valore en una escala del 1 al 5, siendo 1 la valoración más baja y 5 la máxima.

1. Completamente en desacuerdo
2. En desacuerdo
3. Ni de acuerdo ni en desacuerdo
4. De acuerdo
5. Muy de acuerdo

Y por último, ¿podría indicar aproximadamente cuánto tiempo ha tardado en rellenar la encuesta?

1. Más de 5 minutos
2. Aproximadamente 4 minutos

3. Aproximadamente 3 minutos
4. Aproximadamente 2 minutos
5. Menos de 2 minutos.

J. Eventos y trabajos

J.1 Eventos para el fomento de la reutilización de datos abiertos y la edición colaborativa

Para el fomento de la reutilización de los datos abiertos dentro de la comunidad educativa se han desarrollado varios talleres, este tipo de actividades son clave para la difusión y descubrimiento de los conjuntos de datos por parte de la comunidad.

- Este curso 2019-2020 se han realizado dos talleres de edición colaborativa con alumnos de diferentes titulaciones[1]. Este año el tema elegido ha sido dar visibilidad a las mujeres escritoras de habla hispana.

  Alumnos de 4º de Ingeniería Multimedia del itinerario de gestión de contenidos de la Universidad de Alicante. En diciembre los alumnos de la Universidad de Alicante editaron los datos para enlazar y mejorar el enriquecimiento entre Wikidata y `data.cervantesvirtual.com` contribuyendo a la difusión del patrimonio cultural. Editaron y enlazaron datos de algunas obras de escritoras españolas como Carmen de Burgos, Gabriela Mistral, Concepción Arenal, Gloria Fuertes, Fernán Caballero, Gertrudis Gómez de Avellaneda y Emilia Pardo Bazán entre otras.

[1] `http://data.cervantesvirtual.com/blog/2019/12/16/mini-dataton-en-la-ua-y-la-difusion-del-patrimonio-cultural/`

También lo hicieron en otro taller realizado este mes de diciembre de 2019, los alumnos de la asignatura de Nuevas Tecnologías de la Información y comunicación del Turismo del Máster Universitario en Dirección y Planificación del Turismo realizaron un taller de edición colaborativa. Resultados obtenidos después del taller: Presencia de estas autoras en Wikidata y 192 obras dadas de alta.

- #WikiHackatón[2], hackathón basado en el uso y reutilización de datos abiertos a través de las tecnologías de la información para desarrollar soluciones innovadoras e implementar servicios y aplicaciones web que exploten los datos de Wikidata. Se crearon 8 equipos multidisciplinares formados entre 3 y 5 personas que provenían de Ingeniería Informática y de la Ingeniería Multimedia de la UA, con un total de 39 participantes y 8 invitados.

- Wikimedia España[3] y la Fundación Biblioteca Virtual Miguel de Cervantes,[4] en colaboración con el Centro de Competencia IMPACT,[5] organizaron #DatatonCervantes,[6] en Medialab-Prado, Madrid, el sábado, 24 de febrero de 2018. En este evento se consiguió enlazar 1189 obras entre Biblioteca Virtual Miguel de Cervantes (BVMC) y Wikidata.

[2] https://eps.ua.es/es/ingenieria-multimedia/gestioncontenidos/wikihackathon/wikihackaton-2018.html

[3] https://www.wikimedia.es/

[4] https://fundacion.cervantesvirtual.com/

[5] https://www.digitisation.eu/

[6] https://www.wikidata.org/wiki/Wikidata:Events/Madrid/2018-02-24

Hora	Lugar	Actividad	Ponentes
10:30-10:45	Minilab B Planta primera Medialab-Prado	**Bienvenida y presentaciones**	Centro de Competencia IMPACT Fundación Biblioteca Virtual Miguel de Cervantes Wikimedia España
10:45-11:45		**¿Qué es Wikidata?**	David Abián Gustavo Candela Pilar Escobar
11:45-13:00		**Taller de datos literarios**	Actividad colaborativa
13:00-14:00	Cantina Planta baja Medialab-Prado	Descanso y *brunch* conjunto	
14:00-14:30	Minilab B Planta primera Medialab-Prado	**Taller de datos literarios**	Actividad colaborativa
14:30-15:00		**Despedida y cierre**	Centro de Competencia IMPACT Fundación Biblioteca Virtual Miguel de Cervantes Wikimedia España

Figura J.1: Programa del evento DatatonCervantes.

J.2 Dirección de Trabajos Final de Grado orientados a LOD

A continuación se listan algunos de los Trabajos Final de Grado que han sido codirigidos por la autora de esta tesis dentro del contexto de LOD.

- Análisis y visualización de autoras de la Biblioteca Virtual Miguel de Cervantes. En el que se ha creado un sistema de Business Intelligence con Power BI , con un conjunto de cuadros

de mandos que facilitan la visualización, detección de patrones y la extracción de conclusiones sobre las autoras en la Biblioteca Virtual Miguel de Cervantes, además, estos datos se han enriquecido con Wikidata a través de consultas SPARQL.

- Análisis e implementación de una interfaz para visualización de un repositorio RDF[7].
- Análisis e implementación de *Experience API* para recursos online.

[7] `http://hdl.handle.net/10045/96987`