Graduado en Ingeniería Informática

Universidad Politécnica de Madrid

Escuela Técnica Superior de
Ingenieros Informáticos

TRABAJO FIN DE GRADO

Soporte tecnológico para la creación y publicación de vocabularios de datos abiertos

Autor: Javier Pérez Gonzalo

Director: Oscar Corcho García

MADRID, JUNIO 2018

Agradecimientos

Este trabajo ha sido una importante fuente de aprendizaje para mi en ciertos aspectos pero, también ha supuesto tiempo y esfuerzo y es por ello por lo que sin la ayuda de ciertas personas hubiera sido difícil llevarlo a cabo.

Este trabajo esta dedicado a toda mi familia y amigos, los cuales creyeron en mi y siempre pensaron que lograría alcanzar esta meta. Si bien es cierto que no soy un estudiante de notas brillantes, siempre me gusta esforzarme al máximo en las cosas por las que siento interés y me motivan. Espero que este trabajo no sea el final, sino el comienzo de otra etapa llena de inquietudes, de preguntas sin respuesta y de conocimiento.

En especial, quería dar las gracias a mi pareja, Lucía, la cual me ha apoyado no solo durante la realización de este trabajo, sino a lo largo de toda la carrera. Ella ha hecho que quiera ser mejor en muchos sentidos y nunca deje de esforzarme por lograr lo que me proponga, y por ello, le dedico estas líneas.

Por último, quería dar las gracias a mi tutor, Oscar Corcho, siempre dispuesto a ayudar con los problemas que se plantean y proponer alternativas. También al equipo de OnToology por su tiempo; por aportar algunas de las ideas sin las cuales este proyecto no habría sido lo mismo.

En definitiva, gracias por haberme dado la oportunidad de aprender otra vez más.

Madrid, Junio de 2018

Abstract

This final degree project is about the creation and configuration of tools and services needed to provide technological support for the creation and publication of open data vocabularies. In the last few years, a series of initiatives like this have emerged with the aim of creating and publishing vocabularies that define open data sets.

This process has been carried out from repositories that the Spanish data network and Linked Data (OpenCityData) maintains in its GitHub account. These vocabularies need to be updated to be useful and that is why version control over them is an important operation of the process. Therefore, in this project, problems related with the creation and publication of vocabularies will be identified along the whole lifecycle and several solutions will be proposed to solve them. These solutions will be oriented to the maintenance of the vocabularies, prioritizing in the automation and the ease of the operations that are carried out at the different steps of the creation and publication process.

Also, this project can be used as a guide to anyone who wants to know the process, the mistakes that can be found and the possible solutions and tools to consider.

Resumen

El presente Trabajo Fin de Grado consiste en la creación y configuración de las herramientas y servicios necesarios para la prestación de soporte tecnológico para la creación y publicación de vocabularios de datos abiertos. En los últimos años, han surgido una serie de iniciativas como esta con el objetivo de crear y publicar vocabularios que definen conjuntos de datos abiertos.

Este proceso se ha llevado a cabo a partir de los repositorios que la red española de datos abiertos y Linked Data (OpenCityData) mantiene en la plataforma GitHub. Estos vocabularios necesitan ser actualizados para ser de utilidad y es por ello por lo que el control de versiones sobre los mismos es una operación importante del proceso. Por tanto, en el presente trabajo, se identificarán problemas relacionados con la creación y publicación de vocabularios a lo largo de todo su ciclo de vida y se propondrán diversas soluciones destinadas a solventarlos. Estas soluciones estarán orientadas al mantenimiento de los mismos, priorizando en la automatización y la facilidad de las operaciones que se llevan a cabo las diferentes etapas del proceso de creación y publicación.

Además, este trabajo podrá usarse a modo de guía para cualquiera que quiera conocer el proceso, los errores con los que puede encontrarse, así como las posibles soluciones y herramientas a tener en cuenta.

Índice General

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

Hoy en día la tecnología avanza a un ritmo acelerado y cada vez es más común oír conceptos emergentes como: gobierno abierto, ciudad inteligente, y datos abiertos. Es cierto que cada vez más se recogen grandes conjuntos de datos, cuyo catálogo abarca diferentes temáticas, sectores y aspectos, que se generan durante la actividad diaria que se desarrolla dentro de estas ciudades. Los gobiernos abiertos deben hacer un ejercicio de transparencia y de aplicación de los derechos fundamentales propios de los estados modernos. Como consecuencia de esto, surge la necesidad de creación de portales de datos abiertos a disposición de la ciudadanía para que ejerza su derecho tal como se expone el artículo 105.b de la constitución [1].

Aunque en la actualidad, ciudades como Barcelona, Santander, Valencia y Madrid cuentan con portales de datos abiertos, no es hasta 2009 cuando *datos.gob*, la iniciativa de datos abiertos del gobierno de España comienza a promocionar la cultura de la apertura de información en España. Su objetivo es crear las condiciones para el desarrollo del mercado de reutilización de la información del sector público, así como, para dar apoyo a las unidades administrativas, en las unidades técnicas y administrativas necesarias para que publiquen de acuerdo con la legislación vigente y de la forma más amigable para su reutilización la información que recogen [2].

Uno de los problemas a la hora de gestionar, editar, crear y publicar vocabularios de conjuntos de datos abiertos es que no existe un estándar unificado para realizar estas actividades y únicamente se siguen guías de buenas prácticas. En ellas se describen las pautas y procesos a seguir para poder definir un proyecto personalizado de los diferentes procesos de gestión del modelo de apertura de datos pero, es decisión de cada uno implantar o no estas pautas por lo que encontramos portales de datos abiertos en los que

algunos de los *datasets* que proveen no poseen la calidad esperada o no estar enlazados con otros conjuntos de datos para poder ser utilizados correctamente.

Por ello, la Federación Española de Municipios y Provincias (FEMP) [3], mediante su Grupo de Datos Abiertos, creó una guía estratégica para la puesta en marcha de datos abiertos [4], con el objetivo de que cualquier entidad pueda implantar una política que favorezca el gobierno abierto y sirva estos datos a la ciudadanía. En esta guía se detalla el proceso para la apertura y reutilización de los mismos, la cual se basa y resume en los párrafos siguientes, mediante la elaboración de una hoja de ruta que cada administración deberá crear basándose en un plan estratégico general, donde han de describirse las políticas de Gobierno Abierto. Estas hojas de rutas propias han de llevarse a cabo teniendo en cuenta la definición de un plan estratégico y tecnológico, un modelo de datos, una medición en la reutilización y por último una propuesta sobre como llevar a cabo su divulgación.

En cuanto al proceso de apertura de datos abiertos, la guía recomienda un modelo de cómo llevarlo a cabo, identificando diferentes pasos o etapas. En primer lugar, se tendría que llevar a cabo una fase de identificación de los conjuntos de datos que queremos abrir al público, los cuales estarán disponibles en un Portal de Datos Abiertos. La identificación y elección de los conjuntos depende de muchos factores, entre ellos a quién se dirigen (departamentos municipales, ciudadanos, empresas o ciertos grupos de interés). Otra importante etapa del proceso es la priorización. Se debe determinar que conjuntos de datos son más prioritarios a la hora de ser publicados, ya sea por su importancia, demanda o a quién van dirigidos. Para ello se acordará un orden de incorporación o publicación programados. Una vez realizadas estas dos etapas, es necesario realizar una conceptualización del recurso. Esta fase del proceso comprende diferentes tareas. Habrá que identificar las fuentes de los datos, teniendo todas ellas presentes, así como también definir los campos que constituirán el recurso en sí mismo. Además es importante en este punto reunir y aplicar unas buenas prácticas a la hora de definirlo ya que nos evitará errores a posteriori cuando se genere el recurso. También es importante, en la conceptualización del recurso, determinar los formatos en los que el

conjunto de datos a publicar estará disponible así como la frecuencia con la que se deberá realizar la actualización del conjunto. Una vez se ha realizado la conceptualización del recurso, el siguiente paso es la generación del mismo. Se llevará a cabo la generación del fichero, las APIs dinámicas, etc. Además, se creará un proceso de publicación en el que se establezca si es automático o no, su URL, los servidores en los que estará disponible, manuales de uso, etc. Posteriormente, los conjuntos de datos generados se publicarán en un Portal de Datos Abiertos para su exposición al público y se darán de alta en dicho portal. A partir de este momento, se llevará a cabo durante todo el ciclo de vida un mantenimiento del BackOffice para mantener en correcto funcionamiento estos conjuntos de datos.

Dentro de los distintos tipos de modos de apertura de un conjunto de datos abiertos en un Portal, se pueden encontrar un primero basado en la disponibilidad inmediata o a petición de usuario (reactivo). Tiene varias ventajas como el bajo coste y la capacidad para satisfacer directamente una demanda, pero suele tener una utilidad desconocida, los conjuntos de datos no suelen tener un contexto claro y si la solicitud supone un elevado coste de publicación puede dañar la imagen de la iniciativa. El segundo modo se basa en la coordinación y planificación entre departamentos y reutilizadores (proactivo). Permite una mejor planificación y publicación de los datos pudiendo priorizar y otorgando información sobre futuras adiciones. Además, incrementa el número de departamentos involucrados en el proyecto. En cuanto a sus desventajas, es necesario establecer compromisos entre las partes y asignar los recursos de manera adecuada para cumplir con dichos compromisos de manera coherente. Por último se encuentra el modo basado en la coordinación y planificación entre departamentos y reutilizadores (proactivo). En el se refuerza la cultura de apertura de datos abiertos creando una imagen sólida de la corporación municipal donde está presente una planificación entre todas las partes.

A la hora de mantener y actualizar los datos, debemos hacer uso de las herramientas de las que dispongamos con el fin de disponer de una organización coherente de los conjuntos de datos. Desde la guía de puesta en marcha de datos abiertos, se recomienda disponer de herramientas para la limpieza de los datos del conjunto,

detectando errores, campos conjuntos, disjuntos, fuera de rango, etc. Además, aconsejan utilizar una herramienta visualizadora que permita crear representaciones gráficas para facilitar el entendimiento de la información publicada. Es fundamental la documentación bien elaborada y coherente con los datos, favoreciendo la reutilización y aportando valor al ciudadano.

Favoreciendo la filosofía de gobierno abierto, la estrategia para la puesta en marcha de conjuntos de datos ha de usar una plataforma en la que se pongan a disposición de los ciudadanos dichos conjuntos de datos. En los últimos años se ha producido una evolución en las plataformas que se usan para publicar los datos pasando de ser un mero conjunto de enlaces a los datos en crudo a incorporar nuevas funcionalidades: capacidades de búsqueda avanzadas sobre los conjuntos de datos, el uso de herramientas como las antes mencionadas para la visualización de los mismos, capacidad para la recopilación automatizada y actualización de datos, etc. A la hora de elegir una plataforma se recomienda abogar por las facilidades que ofrecen a la hora de mantener una filosofía de reutilización y sobre todo por aquellas plataformas que proporcionen además de una API, una documentación bien elaborada para adquirir correctamente las habilidades necesarias para su uso.

Una parte importante del proceso es cómo se organizan y gestionan los datos y en esta guía también se proporciona una serie de criterios a considerar. Es de vital importancia que los datos sean únicos, que no estén repetidos salvo condiciones excepcionales de seguridad. Tienen que ser compartidos, estar disponibles para todos los que forman parte de una organización y toda aquella persona que tenga interés en los mismos. Por supuesto, los datos tendrán que ser accesibles y abiertos y se deberá favorecer el uso de formatos estándar de uso abierto. Además, siempre que sea posible, se recomienda que los datos estén descritos semánticamente y dicha información se almacene con los mismos. Así, aportarán conocimiento sobre cómo se estructuran los datos y el contexto en el que se encuadran [4].

Por último, unas partes no menos importantes de una iniciativa de datos abiertos son los sistemas de medición y los planes de formación y difusión. Los sistemas de medición permiten ofrecer datos abiertos de acuerdo a la demanda existente y monitorizarlos mediante métricas con el fin de evaluar si se cumplen los compromisos establecidos, mediante el uso de diversos indicadores. Existen multitud de indicadores diferentes entre sí, pudiendo tener en cuenta por ejemplo, los siguientes: número de descargas totales y conjuntos de datos disponibles, número de aplicaciones desarrolladas usando estos datos, etc. En cuanto al plan de formación y difusión, ya que es necesario disponer de personas cualificadas para gestionar datos y desarrollar servicios para el acceso a los mismos, se deben identificar las características de estas personas y de los usuarios que van a utilizar los datos, definiéndose una estrategia correcta para ello. Se elaborarán formaciones para el personal técnico que trabaja con los datos con contenidos tanto teóricos como prácticos. En última instancia, no hay que olvidarse de la formación de la ciudanía, en la que se han propuesto últimamente proyectos colaborativos, y de los reutilizadores con el objetivo de que se desarrollen aplicaciones que hagan uso de datos abiertos [4].

Una vez realizado este análisis de las diferentes estrategias, métodos y buenas prácticas de la Guía estratégica para la puesta en marcha de conjuntos de datos abiertos y bajo el marco de este análisis; este trabajado parte de conjuntos de datos identificados y clasificados en base a su prioridad. Muchos de estos conjuntos ya se encuentran conceptualizados, identificados sus niveles, campos del recurso y fuentes de dichos datos. Otros, requieren la realización de esta conceptualización. Por otra parte, algunos de los recursos ya se encuentran generados y disponen de un proceso de publicación manual.

Como se ha podido observar, este proceso es largo y complejo, por lo que en esta memoria se abordaran los elementos relacionados con la creación y configuración de las herramientas y servicios necesarios para la prestación de soporte tecnológico en las etapas de creación y publicación de vocabularios de datos abiertos. La memoria de este trabajo primero introducirá una serie de conceptos clave para comprender el tema que se desarrollará, para después, desarrollar una serie de posibles soluciones a los problemas

encontrados a lo largo de estas etapas. Para terminar, la memoria finalizará con un conjunto de conclusiones y resultados de todo el proceso.

Capítulo 2

Estudio del dominio

En este capítulo se presentarán los conceptos fundamentales que se van a emplear en la realización del presente trabajo fin de grado y que aparecerán de forma reiterada a lo largo del mismo. Para una mejor categorización, se han subdividido en dos apartados, distinguiendo entre fundamentos y tecnologías básicas.

En el primer y segundo apartado, fundamentos conceptuales y tecnológicos y datos y gobierno abierto, se introducirán nociones tales como ontología, muy importante por ser uno de los conceptos clave que uno debe conocer para entender correctamente este trabajo y como se relaciona con la creación y publicación de vocabularios. Se hablará también de RDF, uno de los formatos utilizados a la hora de describir conjuntos de datos abiertos y que también se relaciona de gran manera con la publicación de los mismos. Se introducirán también RDF Schema y OWL, lenguajes que permiten implementar ontologías. Por otro lado, se definirán términos como portal de datos y gobierno abierto.

En el tercer apartado, tecnologías básicas, se centrará en introducir las tecnologías que usaremos a lo largo de todo el desarrollo del trabajo y que están enfocadas en el mismo. Para ello se anunciarán conceptos como Ontoology, herramienta principal que ha permitido el desarrollo de este trabajo y Github, plataforma encargada de soporte, control de versiones y flujo de cambios.

2.1 Fundamentos conceptuales y tecnológicos

2.1.1 Ontología

En el campo de la web semántica, una ontología define los conceptos y relaciones (términos) usados para describir y representar un área de interés [5]. Se usa para clasificar los términos de una aplicación en particular, caracterizar las posibles relaciones y definir

posibles restricciones en el uso de los términos. Su uso en la web semántica se debe a la utilidad que aporta a la hora de integrar datos cuando existe ambigüedad en los términos de diferentes conjuntos de datos (datasets). Además, también se suele utilizar para organizar conocimiento pudiendo aprovecharse del poder de los datos enlazados. Normalmente, se utiliza el término ontología para colecciones de datos formales y complejos y en cambio se usa el término vocabulario cuando no es necesario un formalismo estricto. Las ontologías suelen incluir comúnmente los siguientes elementos:

- **Clases:** Representan algo que existe en el mundo real y pueden ser conjuntos que comparten una estructura común o forma de comportamiento, ya sean entidades, conceptos, objetos, etc.
- **Propiedades:** Son aspectos, propiedades, rasgos, características o parámetros que los conceptos y las clases poseen o pueden tener, definiéndolas y describiéndolas.
- **Relaciones:** Son las diferentes formas en las que las clases y los individuos (instancias de objetos) pueden estar relacionados unos con otros.

Existen diferentes tipos de ontologías como las de dominio, generales, terminológicas, etc. A la hora de generar y usar las ontologías es preciso codificarlas. Uno de los lenguajes más utilizados hoy en día es OWL ya que está diseñado para aplicaciones que procesan información o la presentan. Además, soporta XML y RDF, siendo este último es el que se utiliza para modelar los elementos de la ontología como recursos, identificándolos por su propia URI. Una de las herramientas que pueden utilizarse para desarrollar ontologías es Protegé [6], un editor desarrollado por la Universidad de Stanford que además, es de código abierto.

2.1.2 RDF

Resource Description Framework (RDF) es un lenguaje estándar del W3C para el intercambio de datos en la Web [7]. Tiene características que facilitan la unión de los datos aun cuando los esquemas que los definen son diferentes. Usando este simple

modelo, permite tanto a datos estructurados como semi estructurados unirse, ser expuestos y compartidos entre diferentes aplicaciones.

Un fichero RDF está formado por tripletas RDF. Una tripleta es una composición de tres elementos relacionados entre sí: sujeto, predicado y objeto. En las tripletas, los sujetos y los objetos se denominan nodos y al predicado también se le da el nombre de propiedad. El sujeto y el predicado siempre son referencias a URIs, en cambio, el objeto puede ser un literal o ser una referencia a una URI, formando una tripleta más compleja. Por ejemplo, la frase "RDF es un framework de descripción de recursos" podría representarse en una tripleta RDF de la siguiente manera:

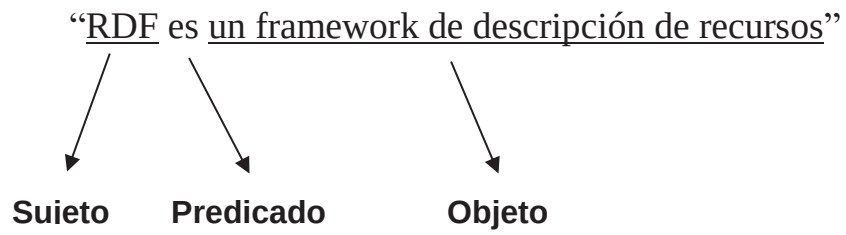


Ilustración 1. Ejemplo RDF

Los datos modelados son válidos hasta que dicha información deje de ser válida o cambie en el futuro y este es un proceso que complejo que requiere tiempo, por tanto, es necesario realizar siempre un primer paso de estudio del dominio que se quiere representar para desarrollar un modelo que sea persistente en el tiempo y sea capaz de beneficiarse del uso de datos enlazados.

A la hora de trabajar con ficheros RDF hay que tener en cuenta que pueden encontrarse en distintos formatos como puede ser RDF/XML, Turtle, N3 o JSON-LD. Estos ofrecen distintas sintaxis con las que trabajar para la especificación de tripletas.

Por último señalar que, una vez definido el término RDF y tripleta, un conjunto de tripletas RDF se define como grafo y a su vez, un conjunto de grafos forma lo que se

denomina un conjunto de datos o *dataset*. Este término, no debe confundirse con el de conjunto de datos abiertos que se presentará en la sección 2.2.1.

2.1.3 RDF Schema

El RDF Schema [8] provee un vocabulario de modelado de datos RDF. Es una extensión semántica de RDF en la que proporciona mecanismos para describir grupos de recursos relacionados así como las relaciones entre dichos recursos, proveyendo los elementos necesarios para la descripción de vocabularios y ontologías. RDF Schema está escrito en RDF usando términos para determinar las características de otros recursos tales como los dominios, los rangos y las propiedades. En él, se describen las propiedades en términos de clases de recursos a las cuales se aplican. Usando estos mecanismos, es sencillo definir propiedades adicionales dentro del dominio o el rango de una clase sin necesidad de redefinirla.

Los recursos pueden ser divididos en grupos denominados clases y los miembros de una clase, se conocen como instancias de dicha clase. Las clases son ellas mismas recursos y a menudo se identifican por IRIs y pueden ser descritos usando propiedades RDF. RDF distingue entre clases y conjuntos de instancias. Asociado con cada clase hay un set, llamado extensión de clase de la clase, el cual es un conjunto de instancias de la clase. Dos clases pueden tener el mismo conjunto de instancias pero ser de diferentes clases. Una clase puede ser un miembro de su misma extensión de clase y puede ser una instancia de si misma.

En cuanto a propiedades, se describe una propiedad RDF como una relación entre recursos sujeto y recursos objetos. La especificación también define el concepto de subpropiedad que puede ser usado para declarar que una propiedad es una subpropiedad de otra. Si una propiedad P1 es subpropiedad de otra P, entonces, todos los pares de recursos que están relacionados con P1 también se relacionan con P. Al contrario, el termino superpropiedad es a menudo usado como la inversa de la subpropiedad. Si una propiedad P1 es superpropiedad de una propiedad P, entonces todos los pares de recursos que están relacionados con P están también relacionados con P1.

2.1.4 OWL

El Lenguaje de Ontologías Web (OWL) [9] está diseñado para usarse para procesar el contenido de la información en lugar de únicamente representarla. OWL facilita un mejor mecanismo de interpretabilidad de contenido Web que los mecanismos admitidos por RDF y RDF Schema, proporcionando un vocabulario adicional junto con la semántica formal.

OWL puede ser usado para representar explícitamente el significado de términos en vocabularios y las relaciones entre esos términos. Esta representación de términos y sus interrelaciones se denomina ontología. OWL tiene mayor capacidad para expresar significado y semántica que RDF y RDF-S por lo que, OWL va más allá de estos lenguajes en su capacidad para representar contenido interpretable por un ordenador para la Web.

Por último, mencionar que OWL proporciona tres lenguajes, cada uno de ellos con un nivel de expresividad mayor que el anterior: OWL Lite, OWL DL y OWL Full.

2.1.5 Datos enlazados (Linked Data)

El término Linked Data hace referencia a una Web de Datos. Una Web de Datos aprovecha una colección de tecnologías de la Web Semántica tales como RDF, OWL, SPARQL entre otras para proveer un entorno en el cual las aplicaciones pueden realizar consultas sobre dichos datos, inferir datos a partir de vocabularios, etc.

Para poder llevar a cabo esta Web de Datos es necesario tener una gran cantidad de datos disponibles en la Web que sigan un formato estándar (RDF), alcanzable y mantenido mediante herramientas de la Web Semántica. Además, es necesario también que las relaciones entre los datos se establezcan de una manera correcta. Para ello, se siguen principios como el uso de URIs como nombres para las entidades, establecer *endpoints* donde poder realizar las consultas de los datos y crear relaciones entre los datos que aporten información útil. Esta colección de *datasets* interrelacionados es lo que denominamos datos enlazados o Linked Data [10] [11].

2.2 Datos abiertos y gobierno abierto

2.2.1 Datos abiertos (open data)

De acuerdo con la definición hecha por *Open Knowledge Foundation* [12], los datos abiertos son datos que pueden ser usados, reutilizados y redistribuidos libremente por cualquier persona con el único requisito de atribuir y compartir la fuente. Esto quiere decir que en cuanto a la disponibilidad y acceso, los datos deben estar disponibles preferiblemente para su descarga a través de internet. Además, debe estar en un formato que favorezca su modificación. En cuanto a la reutilización y redistribución, los datos se deberán proveer bajo términos que permitan su reutilización y redistribución incluyendo su mezcla o unión con otros conjuntos de datos. Además, todo el mundo debe poder utilizar los datos sin ninguna discriminación, las restricciones como las que previenen el uso comercial, no están permitidas.

2.2.2 Portal de datos abiertos

Es una web dedicada a promover el acceso a los datos de una institución o gobierno e impulsar el desarrollo de herramientas creativas para atraer y servir a la ciudadanía. Estos portales disponen de un catálogo de datos, un conjunto de datos abiertos que una institución pone a disposición de la ciudadanía. Estos datos suelen estar disponibles en varios formatos. Además, para cada *dataset* se muestran una serie de características que ayudan a decidir cual usar basándose en frecuencia de actualización, número de descargas o formato disponible. Normalmente en dichos portales se suele disponer de filtros para limitar ciertos sectores, formatos y frecuencias, además de ordenar según diversos parámetros.

Uno de estos portales que se encuentra disponible es de *datos.gob.es* habilitado por el gobierno de España y en el cual se pueden encontrar conjuntos de datos de diversa índole. Por ejemplo, Madrid cuenta con su propio portal de datos abiertos con datos

relacionados con todo lo referente a la ciudad. Esta se encuentra accesible a través de la página de *datos.madrid.es*

2.2.3 Gobierno abierto

El gobierno abierto tiene como objetivo la apertura del gobierno, que la ciudadanía colabore en la creación y mejora de servicios públicos así como en el robustecimiento de la transparencia y la rendición de cuentas frente al secretismo. También es posible encontrarse con el término *open government*, un gobierno que promueve el libre acceso a los datos permitiendo así el ejercicio de la opinión ciudadana.

En España existen iniciativas como la del portal de datos abiertos del gobierno de España mencionada anteriormente así como también la existencia de varias iniciativas a nivel autonómico.

2.3 Tecnologías básicas para nuestro enfoque

2.3.1 GitHub

GitHub[1] es una plataforma líder en el mundo para el desarrollo de software. Esta acoge la mayor comunidad de desarrolladores del mundo con el fin de descubrir contenido, compartirlo y construir mejor software. Entre sus características, permite crear repositorios tanto públicos como privados en los que se fomenta el desarrollo colaborativo.

Esta plataforma es una web basada en permitir tanto a usuarios como a organizaciones aprender, compartir y trabajar de forma conjunta para construir software, proporcionando además un control de versiones basado en Git [13]. Generalmente se usa para código informático y destaca por ofrecer control de versiones de distribución y la funcionalidad de gestionar el código fuente (provisto por Git). Además, añade más

[1] *https://github.com*

funcionalidades colaborativas como el control de acceso, seguimiento de errores, peticiones de nuevas características y wikis para cada proyecto.

Debido a todas estas características que se ofrecen, todos los archivos con los que se van a trabajar (ontologías, documentación, etc.), se encuentran alojados y disponibles en repositorios pertenecientes a la cuenta que OpenCityData[2] tiene dentro de la plataforma.

2.3.2 OnToology

OnToology[3] es un sistema para automatizar parte del proceso colaborativo de desarrollo de ontologías. Este sistema es capaz de analizar y producir diagramas, así como una completa documentación y validación basándose en errores comunes. Todo ello a partir de un repositorio que contenga al menos un fichero en formato OWL.

Una vez que el repositorio que contiene la ontología este registrado y se produzcan cambios en el mismo, OnToology procederá a realizar entre otro, lo siguiente [14]:

- Generará documentación de la ontología en formato HTML
- Generará diagramas de clases y taxonomía
- Creará una *issue* en GitHub con un resumen de la evaluación y un enlace a un informe de evaluación generado usando OOPS! el cual, es una evaluación de la ontología en la que se muestran errores, todo ello organizado en un documento HTML
- Creará un archivo de configuración para cada ontología pudiendo activar o desactivar la generación de documentación.

[2] *https://github.com/opencitydata*
[3] *http://ontoology.linkeddata.es*

Capítulo 3

Análisis

En este capítulo se procederá a realizar una descripción detallada de las tareas previas que se han llevado a cabo sobre la infraestructura conceptual y tecnológica de partida, necesarias para alcanzar los objetivos finales. También se detallarán los recursos consultados durante todo el desarrollo del trabajo.

3.1 Situación de partida de repositorios de OpenCityData y reorganización

El repositorio de OpenCityData[4] se aloja en la plataforma GitHub y está pensado como una cuenta de desarrollo colaborativo de la red temática española de Open Data para ciudades inteligentes (*Smart Cities*). Esta red [15] que comenzó en 2004 tiene como objetivo facilitar el intercambio de conocimiento y buenas prácticas en el área de los datos abiertos y su aplicación a las ciudades inteligentes, grupos de investigación, administraciones públicas, etc.

En este repositorio se almacenan y desarrollan los materiales relacionados con varios vocabularios que comprenden una variedad de temas entre los que se encuentran el urbanismo, la cultura y el turismo por citar algunos de ellos. En cada uno de ellos, se presenta la URI en la que está publicado, si es que ha alcanzado los requisitos para ser publicado. Estos vocabularios publicados pueden encontrarse en la página *vocab.linkeddata.es*. La idea es que estos vocabularios sigan desarrollándose y se creen nuevas versiones de los mismos.

Muchas de las primeras versiones de estos vocabularios se crearon en el 2014 y se encuentran en el repositorio general sobre datos abiertos, que almacenaba los

[4] *https://github.com/opencitydata*

vocabularios que se definían en el contexto de trabajo de datos abiertos de AENOR [16]. La Asociación Española de Normalización y Certificación, es una entidad líder en España para la certificación, tanto de productos y servicios como de sistemas de gestión. Además, esta entidad también es la encargada del desarrollo y difusión de las normas UNE. Una de estas normas es la norma UNE 178301 Ciudades Inteligentes. Datos abiertos (OPEN DATA), siendo la primera norma que establece un conjunto de requisitos para la reutilización de Datos Abiertos u Open Data elaborados o custodiados por el sector público sirviendo además de referencia a la Administración en la implantación y gestión de proyectos de datos abiertos [17]. Debido a que se decidió crear un repositorio por cada uno de los vocabularios en los que se estaba trabajando, surgen problemas para mantener la trazabilidad de los mismos: cuestiones abiertas, gestión de versiones, prioridades, recursos, etc. Por lo tanto, es necesario realizar una labor de actualización, analizando que vocabularios disponen ya de su propio espacio de trabajado, cuales no y cuales necesitan redefinir sus parámetros.

Algunos de los vocabularios disponen de los datos originales obtenidos del grupo de trabajo OjoAlData100[18] y que se muestran en sus correspondientes repositorios. Este grupo surge con el objetivo de identificar los 100 conjuntos de datos abiertos más relevantes que una administración local debería publicar. Un ejemplo de estos parámetros se muestra a continuación:

Núm	163
Clasificación NTI	Urbanismo e infraestructuras
Clasificación NTI (Descripción)	Saneamiento público, Construcción (infraestructuras, equipamientos públicos)
Conjunto de datos	Callejero (Viales, Numeraciones, Tramero, etc.)
Comentarios	Norma UNE 178301:2015
Valor (De 1 peor a 5 mejor)	5
Transparencia	No

Descripción	Dataset que incluye el callejero de la ciudad
Ejemplo	http://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=b3c41f3cf6a6c410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD
Campos mínimos	Numero de id, denominación de la calle, tipo de calle (calle, avenida, paseo, bulevar, etc.), numeración, geolocalización de cada número, geolocalización de la calle, peatonal o no, barrio o distrito, ¿clasificación económica?, sentido de la circulación
Tamaño de la ciudad	Todas
Tipo ciudad (costa, montaña, capital de provincia, etc.)	De cada uno de los municipios con el menor decalaje en las actualizaciones.
Frec. Actualización mínima	Tiempo real
Histórico relevante	Si

Tabla 1. Ejemplo de parámetros informativos del vocabulario

Por tanto, el primer paso en el análisis realizado consistió en analizar y extraer todo el material y los recursos reutilizables que pueda albergar el repositorio general sobre datos abiertos. Este se encuentra en estado obsoleto y desactualizado por lo que habrá que realizar esta primera tarea poniendo atención a los detalles.

Actualmente, este repositorio alberga los siguientes vocabularios y tan solo algunos de ellos contienen referencias a la versión a la que hace referencia y que está publicada.

Vocabulario	Contiene referencia a versión
turismo/lugar	Si
cultura-ocio/agenda	Si
hacienda	No
sector-publico/organizacion	No
sector-publico/subvencion	No
sector-publico/territorio	No
transporte/trafico	Si
turismo/alojamiento	Si
urbanismo-infraestructuras/callejero	No
Urbanismo-infraestructuras/equipamiento	No

Tabla 2. Lista de vocabularios y referencias a versiones

Una vez se ha realizado esta comprobación, hay que acudir al repositorio de OpenCityData y hay que proceder a realizar una comparación entre cada uno de los vocabularios existentes en el antiguo repositorio y los alojados en este, ahora dispuestos en repositorios independientes para cada uno de ellos. Habrá que crear si no existiera, repositorios adicionales para cada nuevo vocabulario. En caso de que exista, se procederá a comprobar y actualizar la información existente para que no haya ambigüedades entre versiones. A continuación, se muestra una tabla que refleja la estructura del repositorio de OpenCityData:

cultura-ocio	cultura-ocio-salas-ocio
	cultura-ocio-bibliotecas
	cultura-ocio-bibliotecas-prestamos
	cultura-ocio-bibliotecas-catalogo-libros
	cultura-ocio-lugares-interes
	cultura-ocio-agenda-eventos-actividades
	cultura-ocio-museos-galerias-exposiciones

hacienda	hacienda-presupuesto
	hacienda-actividad-inspectora
	hacienda-impuestos-personas-juridicas
	hacienda-impuestos-agentes
	hacienda-licencia-terrazas
	hacienda-ejecución-presupuestaria
sector-publico	sector-publico-servicio
	sector-publico-organismos
	sector-publico-subvenciones-ayudas
	sector-publico-declaraciones-compatibilidades
	sector-publico-agendas
	sector-publico-arrendamientos-alquileres
	sector-publico-puestos-trabajo
	sector-publico-convenios-contratos
	sector-publico-facturas
	sector-publico-registros
	sector-publico-encomiendas-gestion
	sector-publico-normativa-municipal
	sector-publico-territorio
	sector-publico-contrataciones-licitaciones-servicios
transporte	transporte-aparcamiento
	transporte-urbano
	transporte-semaforos
	transporte-restricciones-circulacion
	transporte-bici-carriles
	transporte-aparcamiento-zonas-reguladas
	transporte-bici-datos-publicos
	transporte-parques-vehiculos
	transporte-senializacion-vertical
	transporte-sanciones-multas-trafico
	transporte-senializacion-horizontal
	transporte-trafico-tiempo-real
	transporte-accidentalidad-trafico

turismo	turismo-oficinas-puntos-informacion
	turismo-alojamiento
urbanismo - infraestructuras	urbanismo-infraestructuras-alumbrado-publico
	urbanismo-infraestructuras-callejero
	urbanismo-infraestructuras-cartografía
	urbanismo-infraestructuras-actuaciones-urbanisticas
	urbanismo-infraestructuras-licencias-obra
	urbanismo-infraestructuras-ionspección-edificios
	urbanismo-infraestructuras-quioscos
	urbanismo-infraestructuras-bienes-inmuebles
	urbanismo-infraestructuras-fuentes-agua-potable
	urbanismo-infraestructuras-obras-conservacion-urbana
comercio	comercio-catalogo-empresas
	comercio-censo-locales
	comercio-mercadillos-puestos-municipales
	comercio-empresas-sistema-arbitral-consumo
medio-ambiente	medio-ambiente-calidad-agua
	medio-ambiente-climatologia
	medio-ambiente-areas-verdes
	medio-ambiente-niveles-polinicos
	medio-ambiente-arbolado
	medio-ambiente-puntos-limpios-contenedores
	medio-ambiente-contaminación-acustica
	medio-ambiente-calidad-aire
	sociedad-bienestar-equipamientos-municipales

Tabla 3. Estructura del repositorio de OpenCityData

3.1.1 Resolución de problemas de nombrado

Una vez que tenemos una visión clara de la estructura y los contenidos de ambos, se puede empezar a hablar de realizar una reorganización de estos. A primera vista, encontramos problemas de nombrado de los vocabularios por lo que la equivalencia entre

unos y otros deberá realizarse en base a su contenido. Una vez analizados uno por uno el contenido, estos problemas de equivalencia pueden resolverse siguiendo la siguiente correspondencia propuesta:

Repositorio de OpenCityData	**Repositorio obsoleto**
cultura-ocio-lugares-interes	turismo/lugar
cultura-ocio-agenda-eventos-actividades	cultura-ocio/agenda
hacienda-presupuesto	hacienda
sector-publico-organismos	sector-publico/organizacion
sector-publico-subvenciones-ayudas	sector-publico/subvencion
sector-publico-territorio	sector-publico/territorio
transporte-trafico-tiempo-real	transporte/trafico
turismo-alojamiento	turismo/alojamiento
urbanismo-infraestructuras-callejero	urbanismo-infraestructuras/callejero
sociedad-bienestar-equipamientos-municipales	urbanismo-infraestructuras/equipamiento

Tabla 4. Equivalencias entre los vocabularios de ambos repositorios

De la equivalencia entre repositorios mostrada en la tabla anterior, solo fue necesario la creación de un nuevo repositorio al no ser posible encontrar uno que se asemejara a este, turismo-alojamiento. Además, antes de la reorganización, marcado en naranja en la siguiente tabla, se muestran los vocabularios del repositorio de OpenCityData que poseían el fichero de ontología con extensión OWL y en verde los que lo poseen actualmente debido a que han sido añadidos.

sector-publico-servicio
sector-publico-territorio

urbanismo-infraestructuras-alumbrado-publico
urbanismo-infraestructuras-callejero
medio-ambiente-contaminacion-acustica
medio-ambiente-calidad-aire
sociedad-bienestar-equipamientos-municipales
cultura-ocio-lugares-interes
cultura-ocio-agenda-eventos-actividades
hacienda-presupuesto
sector-publico-organismos
sector-publico-subvenciones-ayudas
transporte-trafico-tiempo-real
turismo-alojamiento

Tabla 5. Vocabularios que disponían del fichero de ontología antes (naranja) y después (verde) de la reorganización

Una vez realizadas estas actualizaciones de repositorios, hay que analizar los ficheros que definen el vocabulario para comprobar que las referencias a las que se dirigen en dichos ficheros sigan siendo válidas, ya que podrían haber cambiado debido a la versión o al cambio del nombre del repositorio. Todos estos cambios se han especificado como issues en GitHub para llevar a cabo un seguimiento de las mismas.

Los siguientes repositorios han sufrido modificaciones debido a estos problemas mencionados. Junto con ellos se adjunta el cambio que ha sido necesario para mantenerlo actualizado. El cambio específico se reporta en GitHub.

Repositorio	**Cambios**
cultura-ocio-lugares-interes	Actualización de las URI del recurso que referencian al html y al fichero de la ontología
cultura-ocio-agenda-eventos-actividades	Actualización de la URI del recurso que referencia al fichero de la ontología

hacienda-presupuesto	Actualización de las URI del recurso que referencian al html y al fichero de la ontología
sector-publico-subvenciones-ayudas	Actualización de las URI del recurso que referencian al html y al fichero de la ontología
sector-publico-territorio	Actualización de la URI del recurso que referencian al html y el cambio de la versión
transporte-trafico-tiempo-real	Actualización de las URI del recurso que referencian al html y al fichero de la ontología
turismo-alojamiento	Actualización de la URI del recurso que referencia al html
urbanismo-infraestructuras-callejero	Actualización de la URI del recurso que referencia a la versión y añadida URI de referencia a la ontología

Tabla 6. Modificaciones en los repositorios

Además, durante el análisis de estos repositorios se han encontrado problemas que afectan a todo el vocabulario. Uno de ellos es sector-publico-organismos, en el que todas las URIs de *datos.gob.es* están desactualizadas y los recursos a los que referencian, no se encuentran disponibles. En el repositorio de urbanismo-infraestructuras-callejero, también encontramos problemas al contener una URI que referencia a un recurso inexistente. En cuento al repositorio de sociedad-bienestar-equipamientos-municipales, hay errores en cuanto a la definición de la ontología y no se puede encontrar una versión publicada que se corresponda. Además, no existe una clara igualdad entre ambas definiciones por lo que se debe realizar un estudio en mayor profundidad para determinar su resultado.

A continuación, se presenta dicho estudio, con el objetivo de determinar cuáles son las propiedades y clases que deben mantenerse y cuáles deben cambiarse o eliminarse. Algunas de ellas llegan a coincidir con el nombrado, marcadas en verde, y en cambio hay otras que pueden llegar a ser adaptadas.

*Se utilizará el prefijo **equip** para <http://vocab.linkeddata.es/ datosabiertos /def/ urbanismo-infraestructuras/equipamiento> y **mun** para <http://vocab.linkeddata.es/ datosabiertos/def/urbanismo-infraestructura/equipamiento-municipal> para mejorar la legibilidad de la tabla.*

equipamientos	**sociedad-bienestar**
Object Properties	**Object Properties**
http://geonames.org/ontology#featureClass	http://schema.org/address
http://geonames.org/ontology#featureCode	http://vocab.linkeddata.es/datosabiertos/def/sector-publico/Servicio#realizacion
http://purl.org/dc/terms/type	**mun:**poseeUn
equip:tipoEquipamiento	**mun**:tipoEquipamiento
equip:tipoAccesibilidad	-
equip:numPlazas	-
Data Properties	**DataProperties**
equip:modoAcceso	http://schema.org/addressLocality
-	http://schema.org/addressRegion
-	http://schema.org/postalCode
-	http://schema.org/streetAddress
-	http://schema.org/url
-	http://schema.org/ContactPoint#email
-	http://schema.org/ContactPoint#faxNumber
-	http://schema.org/ContactPoint#telephone
-	http://schema.org/Thing/ContactPoint#URL
equip:accesible	**mun:**accesible

-	**mun:**descripcion
-	**mun:**horario
-	**mun:**id
-	**mun:**nombre
-	**mun:**tasa
equip:titularidad	**mun:**titular
Classes	**Classes**
http://geonames.org/ontology#Feature	http://schema.org/ContactPoint
http://schema.org/CivicStructure	http://vocab.linkeddata.es/datosabiertos/def/sector-publico/Servicio
http://schema.org/ParkingFacility	**mun:**Equipamiento
equip: Aparcamiento	**mun:**TipoEquipamiento
http://www.w3.org/2004/02/skos/core#Concept	http://www.opengis.net/ont/geosparql
http://www.w3.org/2004/02/skos/core#ConceptScheme	-

Tabla 7. Estudio comparativo de los diferentes componentes en ambos vocabularios

3.2 Implementación de cambios

A la hora de trabajar con la infraestructura tecnológica del proyecto, los medios técnicos y los servicios utilizados para el desarrollo de los vocabularios contenidos en el repositorio de OpenCityData, he hallado una serie de inconvenientes que deben ser subsanados mejorando con esto la infraestructura. El principal inconveniente encontrado es la desactualización de algunos de los vocabularios presentes en el repositorio, algunos de ellos con fechas de última modificación de hace varios años. Sumado a esto, algunos vocabularios contienen issues que no han sido solucionadas y cuyos errores quizá se deban a este motivo.

Los nuevos repositorios que albergan contenido de vocabularios procedentes del antiguo, han sido actualizados. Su fichero de ontología ha sido modificado para referenciar correctamente a las versiones, ubicaciones del fichero de la ontología disponible y documentación. Esto se llevó a cabo en primer lugar como se detalló en el apartado anterior. La documentación a priori no generada por algunos de ellos, ha sido realizada mediante el registro de los vocabularios en la herramienta OnToology, asegurando así que el vocabulario esté provisto de documentación actualizada a medida que se realizan cambios en la ontología. En primer lugar, tras realizar esta centralización de repositorios, tan solo los vocabularios de sector-publico-territorio y urbanismo-infraestructuras-callejero disponían de esta actualización de documentación proporcionada por OnToology. Este último además, estaba generando issues mediante la herramienta de evaluación de ontologías OOPS! proporcionada por OnToology relacionadas con la falta de *disjoint axioms* en la ontología. Un *disjointness axiom* entre dos clases de una ontología declara que un elemento no puede ser instancia de ambas clases [19]. Finalmente se solucionó introduciendo estos axiomas entre las clases de la ontología. A continuación se muestra un ejemplo de la información proporcionada por la evaluación de OOPS! [20]

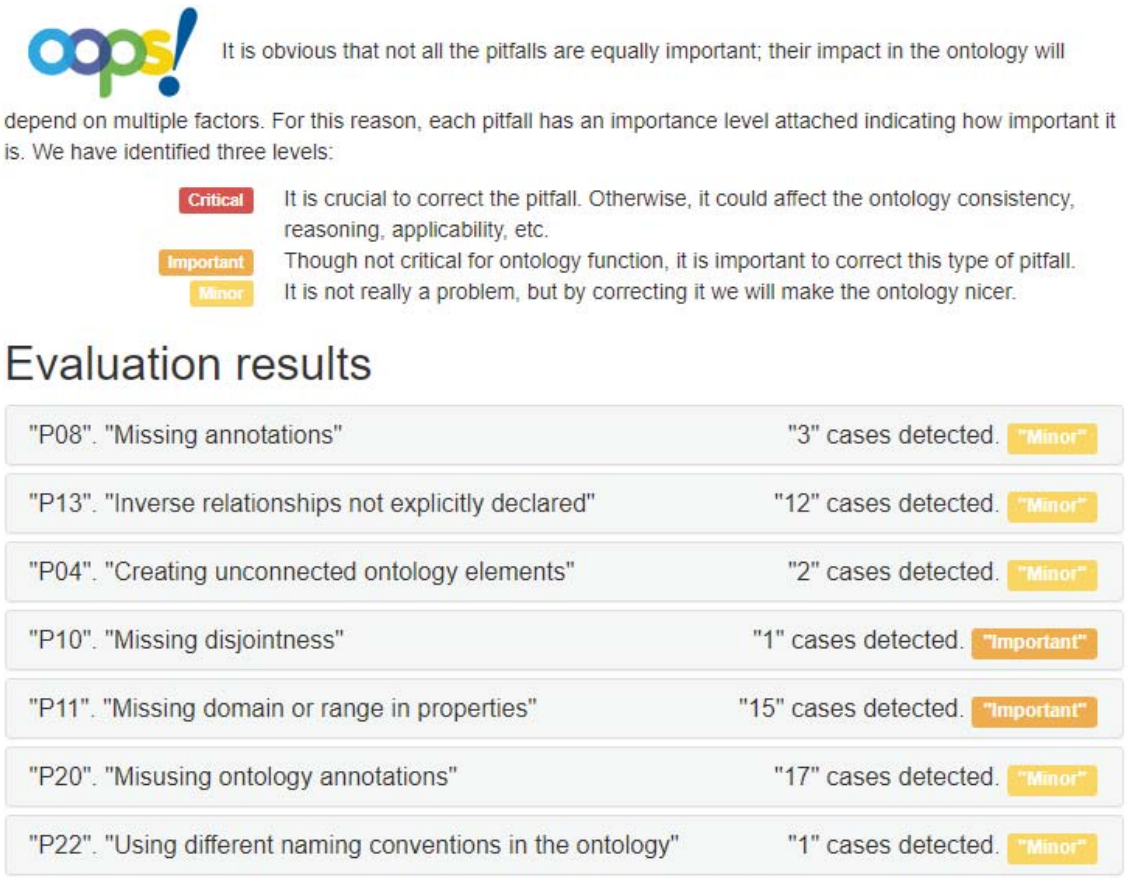


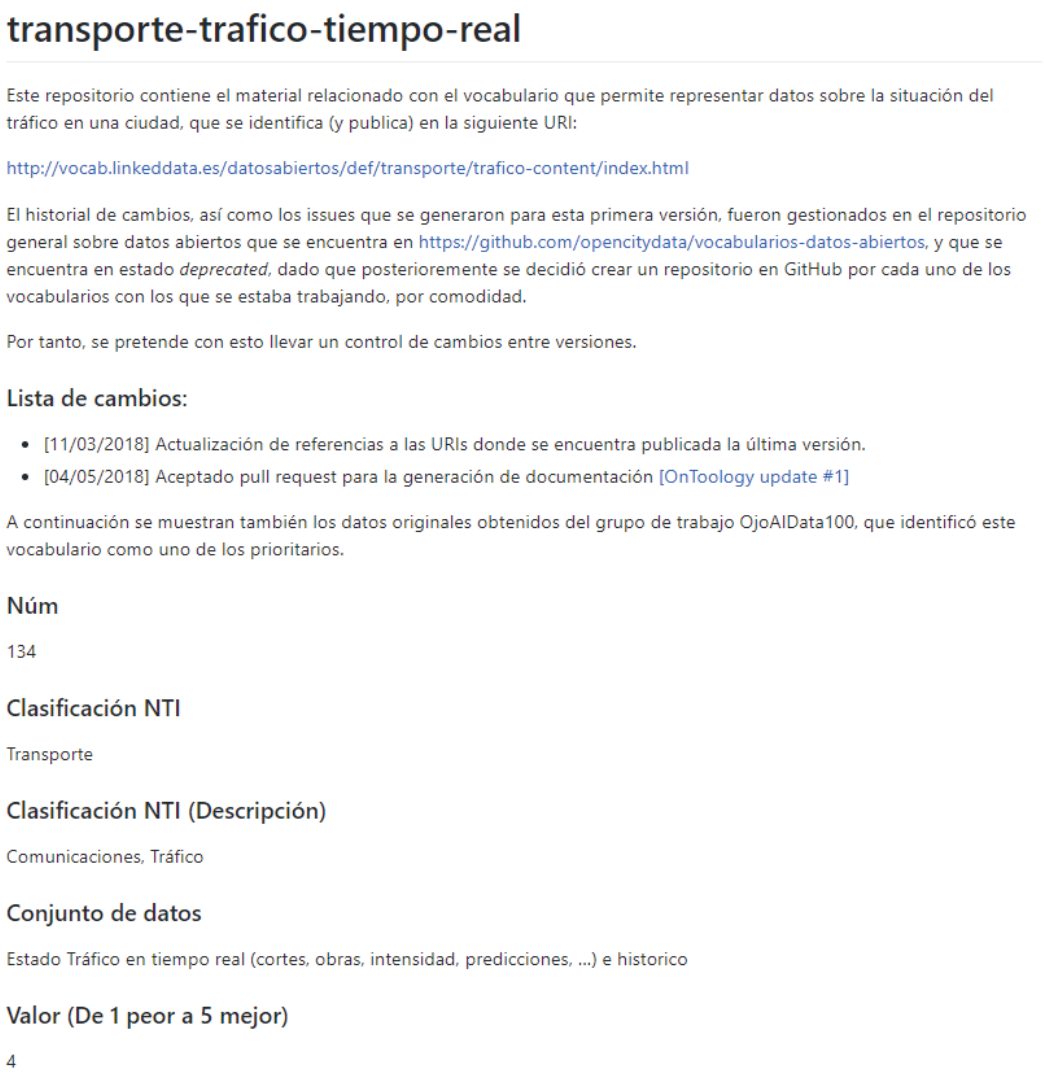
Ilustración 2. Información proporcionada por OOPS!

Como se puede observar, esta herramienta proporciona una evaluación de la ontología, clasificando los problemas encontrados en función de la importancia de los mismos. Dentro de cada apartado se muestra información más concreta del problema que se debe solucionar. En este apartado, la herramienta quizá podría mejorar incluyendo detalles más específicos con respecto al problema. En el caso anterior, podría sugerir cómo corregir los errores, identificando que son las clases de la ontología y no las propiedades, por ejemplo, en las que hay que introducir dichos axiomas. Además, a la hora de registrar los repositorios, debes ser propietario del mismo por lo que podría extenderse este derecho a colaboradores o miembros de organizaciones.

El resto de los vocabularios fueron posteriormente registrados en OnToology, asegurando así que los vocabularios dispongan de documentación actualizada mientras continúe el desarrollo de la ontología. A continuación, se generaron nuevos ficheros

READ.ME para cada uno de los vocabularios en los que se muestra de una forma clara y concisa cual es el propósito del vocabulario, una descripción del mismo, referencias a la documentación, lugar de publicación del vocabulario, así como los datos originales obtenidos del grupo de trabajo OjoAlData100. Un importante añadido a estos ficheros es la lista de cambios. En esta lista se reflejan todas las acciones que han supuesto un cambio en el repositorio de dicho vocabulario tales como issues, pull request aceptados, cambios en el fichero de la ontología, etc. Esta sección es una manera clara y concisa de a simple vista comprobar el estado en el que se encuentra el vocabulario y de navegar entre sus cambios mediante el uso de links a los correspondientes cambios si se da el caso. A continuación se muestra un ejemplo[5].

[5] *https://github.com/opencitydata/transporte-trafico-tiempo-real*



transporte-trafico-tiempo-real

Este repositorio contiene el material relacionado con el vocabulario que permite representar datos sobre la situación del tráfico en una ciudad, que se identifica (y publica) en la siguiente URI:

http://vocab.linkeddata.es/datosabiertos/def/transporte/trafico-content/index.html

El historial de cambios, así como los issues que se generaron para esta primera versión, fueron gestionados en el repositorio general sobre datos abiertos que se encuentra en https://github.com/opencitydata/vocabularios-datos-abiertos, y que se encuentra en estado *deprecated*, dado que posterioremente se decidió crear un repositorio en GitHub por cada uno de los vocabularios con los que se estaba trabajando, por comodidad.

Por tanto, se pretende con esto llevar un control de cambios entre versiones.

Lista de cambios:

- [11/03/2018] Actualización de referencias a las URIs donde se encuentra publicada la última versión.
- [04/05/2018] Aceptado pull request para la generación de documentación [OnToology update #1]

A continuación se muestran también los datos originales obtenidos del grupo de trabajo OjoAIData100, que identificó este vocabulario como uno de los prioritarios.

Núm

134

Clasificación NTI

Transporte

Clasificación NTI (Descripción)

Comunicaciones, Tráfico

Conjunto de datos

Estado Tráfico en tiempo real (cortes, obras, intensidad, predicciones, ...) e historico

Valor (De 1 peor a 5 mejor)

4

Ilustración 3. Ejemplo de información de repositorio

3.3 Generación y publicación

Una vez que se han realizado todos estos pasos, el objetivo último es la generación de la versión de la ontología y su posterior publicación. Debido a todo el intercambio de información entre repositorios, reorganización de los mismos y registro en herramientas como OnToology, mucha de la información que estaba presente y publicada había dejado ser válida o no se estaba generando correctamente. El problema es que al haber realizado cambios en los directorios, las rutas y el nombrado, OnToology no reconocía esta documentación preexistente del vocabulario y generaba un esqueleto vacío de información que había que cumplimentar a posteriori. Tras ubicar y unificar toda la

documentación disponible publicada anteriormente para cada repositorio, se realizo una estructura de directorios clara para que no hubiera problemas a la hora de regenerar la documentación en las veces sucesivas y se procedió a realimentar todos los vocabularios con toda la información disponible recabada.

A la hora de realizar el proceso de reorganización y generación han surgido problemas debido a la estructura que se utiliza para almacenar y desarrollar las ontologías. Conjuntamente con la generación de documentación, se quiere disponer de una generación automática de visualización de los vocabularios publicados que enlace a repositorios de Github, en multi idioma y que el español sea el de por defecto. También sería deseable incluir metadatos de las ontologías y dicha visualización podría ser provista por OnToology. El problema reside en la forma de trabajar de la herramienta, la cual a partir de un repositorio que contenga ficheros de ontología es capaz de generar previsualizaciones de dichos ficheros usando para ello una *GitHub page*. Debido a que OpenCityData estructura cada vocabulario con su fichero de ontología en un repositorio independiente, no existiría una manera a priori de generar una visualización en conjunto de todos los vocabularios, llegando solamente a generar una por cada repositorio existente de manera independiente de los demás.

Como solución a este problema se ha planteado el uso de un repositorio agregador. La idea de este repositorio es agrupar dentro de un repositorio principal, todos los repositorios de vocabularios a partir de los cuales se quiere generar su visualización. El motivo de esta idea es hacer uso de las herramientas que nos proporciona GitHub para ello, mediante la cual cada uno de los repositorios dentro de este principal se verían como submódulos de este, creando una estructura de directorios y a su vez siendo posible desarrollar cada uno de estos directorios por separado sin que se interfiriesen entre sí. Esta idea se puso a prueba sin éxito ya que aunque para la plataforma se trabaje como si de directorios se tratase, es posible que OnToology no sea capaz de leer esta estructura al no tratarse de ficheros propiamente presentes en el repositorio principal y por tanto, no ser capaz de reconocer los ficheros de ontología contenidos en los mismos, haciendo imposible generar la visualización que se desea. La idea sobre esta capacidad de

generación de visualizaciones a partir de repositorios agregados se reportó en el repositorio de GitHub de OnToology con el fin de que se añadiera próximamente y pudiera resolverse este problema[6].

Esta característica se añadió poco antes de la entrega de esta memoria por lo que se decidió incorporarla al proyecto. Por tanto, dentro de OpenCityData se creo un nuevo repositorio que contiene un submódulo [21] por cada uno de los vocabularios a publicar. Cada submódulo es un proyecto (repositorio que contiene una ontología) dentro del principal, por la que se pueden desarrollar de forma separada en cada módulo sin interferir con el principal y manteniendo los *commits* de forma separada, aunque en nuestro caso nos sirve para generar la visualización. También permite actualizar el proyecto principal con los cambios que se hayan realizado de manera independiente en los diferentes submódulos con el objetivo de tenerlo actualizado por lo que es una buena herramienta para la tarea que nos ocupa. Una vez creada la estructura del repositorio, ya se puede registrar en OnToology para que comience a generar la documentación asociada a todos los vocabularios de los submódulos y la previsualización. A continuación se muestra una imagen de la estructura del repositorio[7].

[6] *https://github.com/OnToology/OnToology*
[7] *https://github.com/opencitydata/visualization*

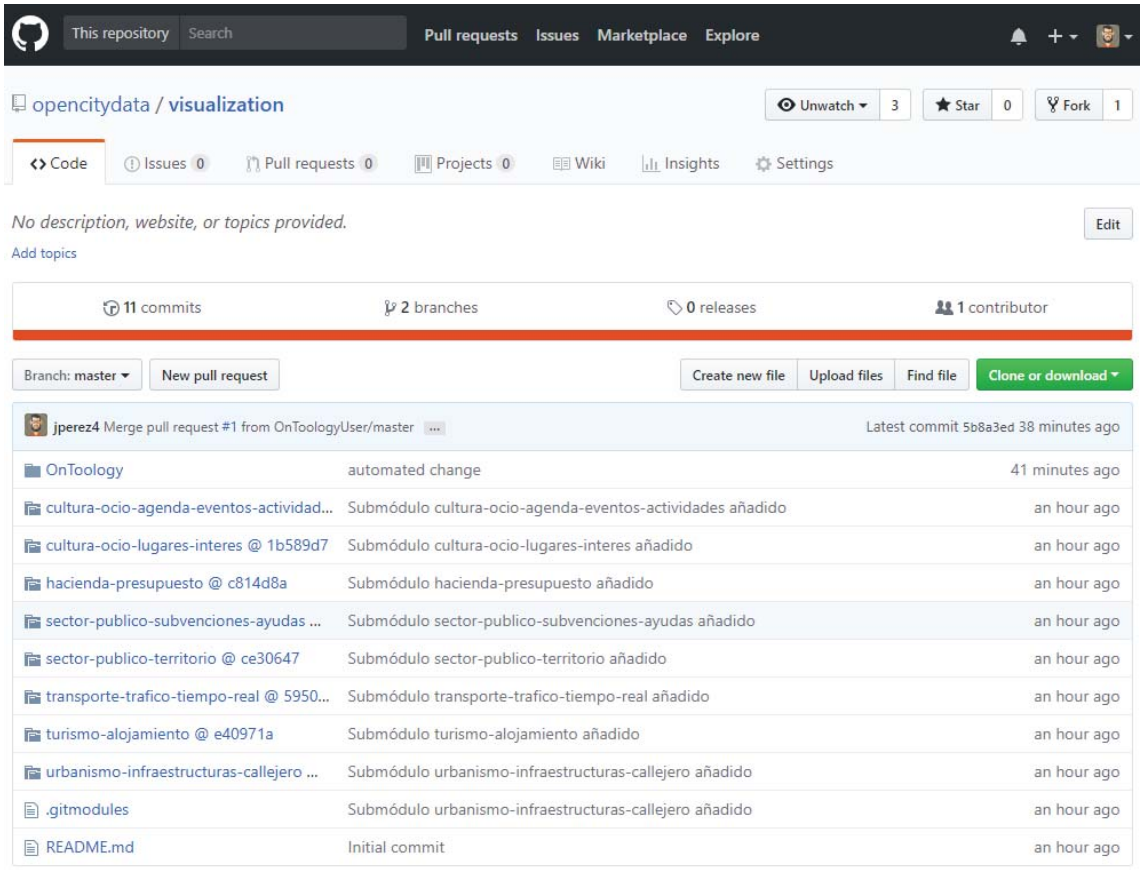


Ilustración 4. Estructura del repositorio de visualización

Una vez que OnToology ha generado la documentación correspondiente a todos los vocabularios, ya se permite generar la visualización. Para ello, OnToology genera una *Github page* dentro del repositorio agregador que contendrá el contenido generado. A continuación se muestra un ejemplo[8]:

[8] *https://opencitydata.github.io/visualization/*

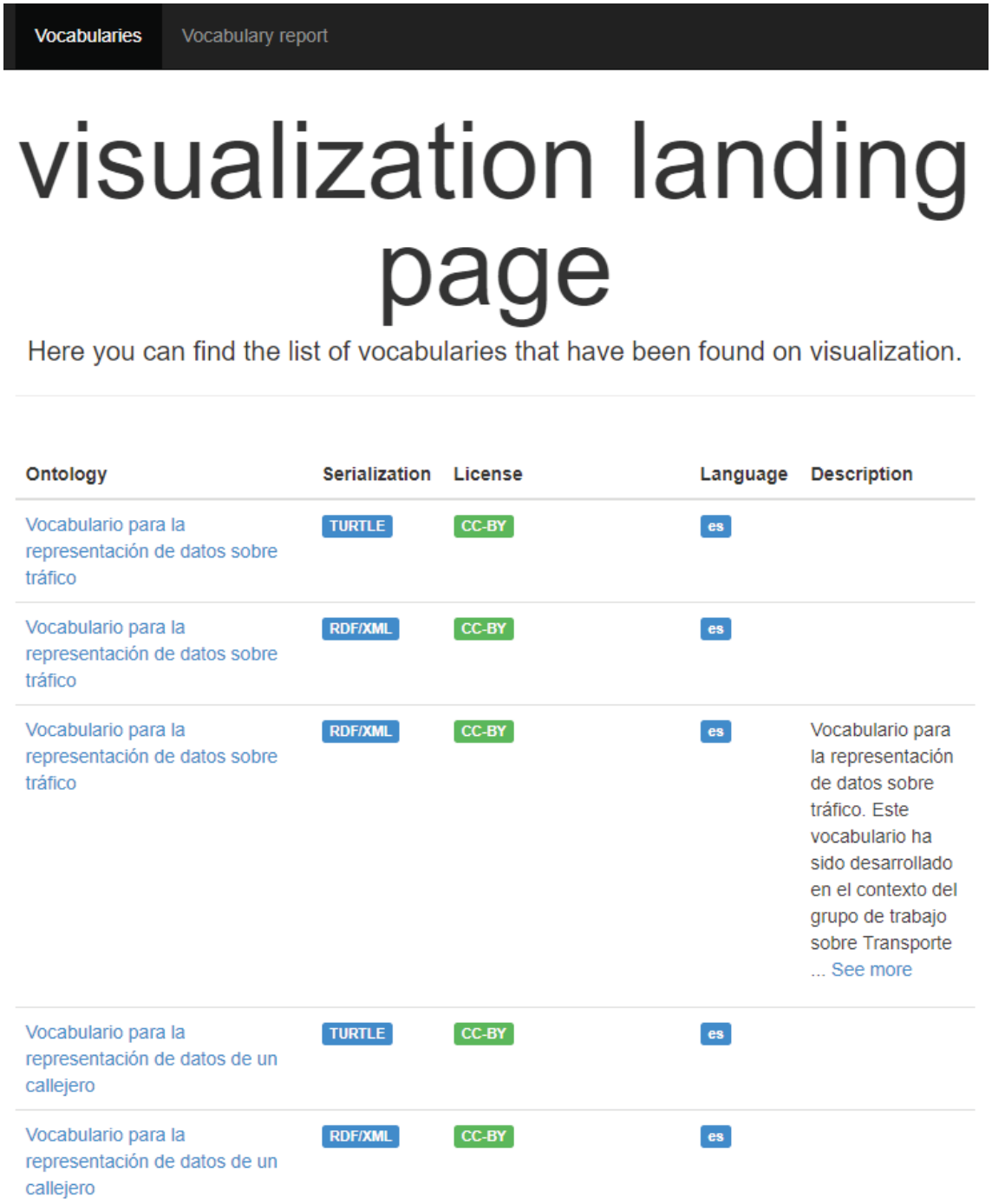


Ilustración 5. Ejemplo de visualización generada

En esta página generada, se pueden observar los vocabularios que están contenido en dicho repositorio agregador y la documentación asociada a los mismos. En el apartado *Ontology*, se dispone de un enlace a la ontología. *Serialization* nos indica en qué formato se encuentra el vocabulario, pudiendo encontrarse en varios distintos. Próximamente será

necesario agregar todas estas serializaciones en un mismo grupo para un mismo vocabulario, permitiendo una mejor visualización. Otros campos que pueden verse son la licencia de uso, el idioma y la descripción. Para este último campo ha sido necesario añadir una etiqueta *description* en cada uno de los vocabularios para que se generara correctamente. Además, se dispone de una pestaña *Vocabulary report* donde se puede comprobar para cada vocabulario, su estatus y los problemas ocasionados. A continuación se muestra un ejemplo[9]:

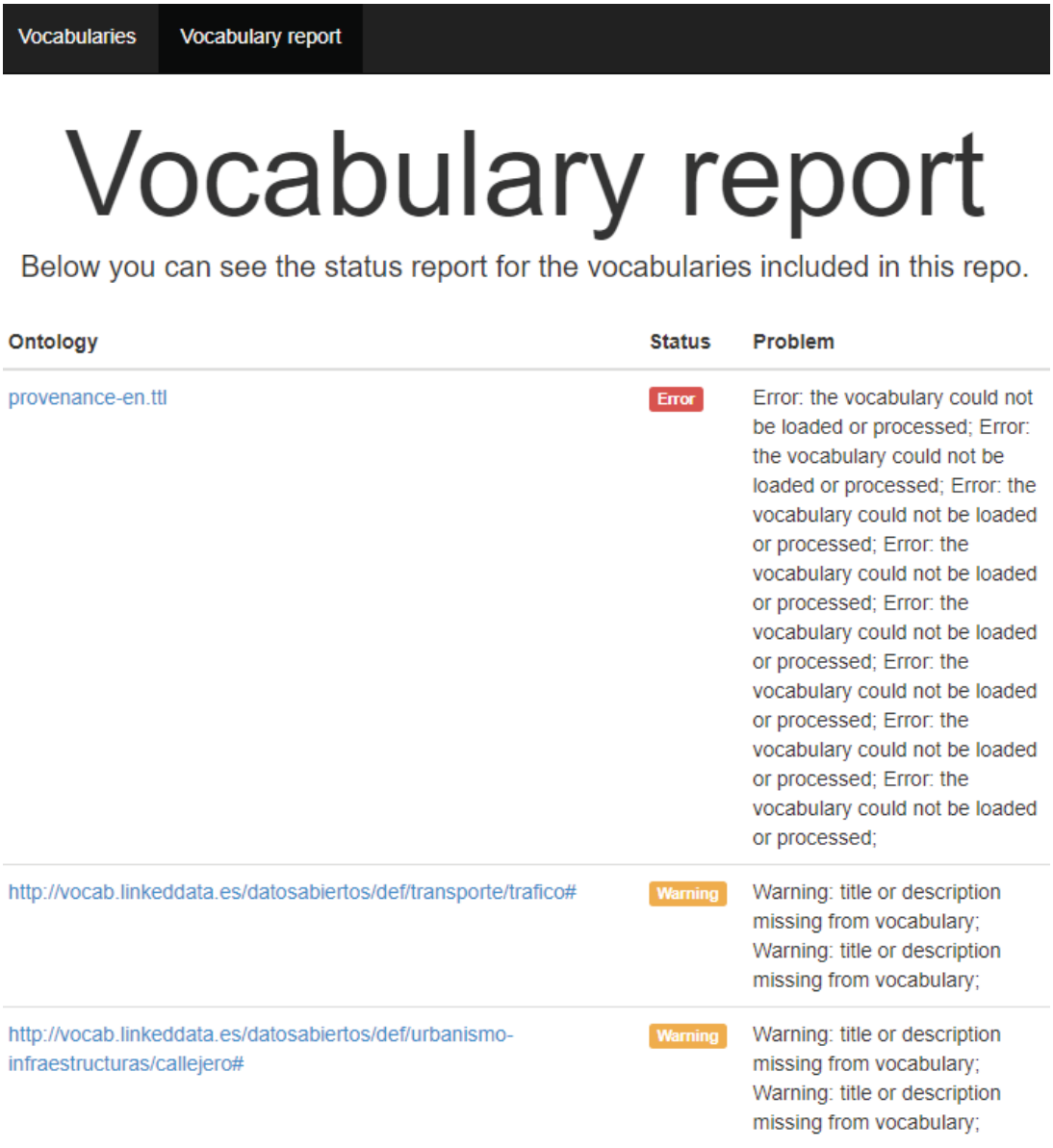


Ilustración 6. Ejemplo de reporte de errores para los vocabularios

[9] *https://opencitydata.github.io/visualization/report.html*

En cuanto a la publicación de los vocabularios se ha llevado una revisión de las propiedades y documentación generada por los mismos antes de realizar este paso. En estas revisiones han solucionado errores que se habían pasado por alto en iteraciones anteriores. Se ha comprobado que todos los vocabularios generen a priori toda la documentación de la que se disponía previamente y se ha actualizado proveyendo información actualizada y adicional a los mismos como la actualización de referencias, nuevos casos de uso, inclusión de diagramas de clases y relaciones, etc. Para realizar este paso es necesario conocer OnToology y los archivos de configuración que se generan junto con la documentación ya que no editar estos archivos supondrá un paso atrás una vez que se modifique la ontología. Este fichero de configuración que puede encontrarse en el directorio principal de la documentación nos permite establecer importantes parámetros de la ontología como el título, el prefijo que se va a usar, la URI, el nombre de la ontología, versiones actuales y anteriores entre otras muchas opciones para personalizar. Este archivo es uno de los primeros que se debe modificar ya que los parámetros que se establecen en el son fijos para el resto de las actualizaciones de la ontología, es decir, cuando la ontología sufra modificaciones, los campos del archivo html que se refieran a parámetros de este archivo de configuración van a ser reemplazados automáticamente. No darse cuenta de esto lleva a tener que realizar más revisiones de las previstas pero es una herramienta útil ya que se realizan todos los reemplazos automáticamente cada vez que se produce un cambio en el archivo de la ontología. Realizada esta configuración, hay que modificar las secciones de la documentación. Aquí OnToology ofrece una estructura muy clara de secciones con sus correspondientes campos, que deben sustituirse por la información que se quiere mostrar. Las secciones de *abstract*, introducción y referencias ofrecen bastante rango de modificación, y se pueden incluir cualquier tipo de elementos ya sean listas, imágenes, *snippets* de código, etc. Sin embargo, las secciones de *crossref* y *overview* que detallas las clases, propiedades y relaciones no ofrecen esta libertad y son menos modificables. Una vez documentadas todas estas secciones, se obtiene la libertad de modificar la ontología y automáticamente se generará una nueva documentación actualizada manteniendo todos estos campos y secciones lo que conlleva a un ahorro de tiempo más que considerable.

Tras realizar estas revisiones de los vocabularios, se puede seguir con el proceso de publicación. Para ello me puse en contacto con el administrador de la web de *vocab.linkeddata.es* y acordamos la subida de toda la documentación asociada con los vocabularios a la plataforma Nextcloud [22], a partir de la cual, el podría acceder a dicha documentación y realizar las acciones necesarias para publicarla. En un primer intento de publicación, hubo problemas con la estructura de la documentación y con contenidos incoherentes. Estos problemas los generaban los vocabularios cultura-ocio-agenda-eventos-actividades y sector-publico-organismos. El primero de ellos contenía una estructura de directorios que no reflejaba correctamente la estructura unificada que se pretendía conseguir con todos los directorios ya que tenía carpetas con archivos pertenecientes a versiones anteriores. Se decidió entonces establecer una reestructuración de todos los repositorios con el fin de eliminar directorios sin utilidad y proveer una estructura mucho más clara e intuitiva con lo que se solucionó este problema. En cuanto a sector-publico-organismos no se pudo resolver los problemas que ocasiona debido a que la ontología mantiene referencias que han dejado de existir y no pueden reemplazarse por lo que el proceso de generación de documentación muestra errores continuados y ficheros inesperados con contenido incongruente, por lo que se ha decidió posteriormente no incluirlo en la publicación. Además hubo problemas generalizados con las plantillas ya que algunas de ellas seguían manteniendo su aspecto de la versión anterior y algunas secciones no reflejaban correctamente los cambios. Esto lleva a pensar sobre la forma de obtener un proceso para publicar de una manera fácil las ontologías con todas sus redirecciones correspondientes y si fuera posible, automatizar todo el proceso.

Actualmente el proceso se lleva a cabo de la siguiente manera. Primero se accede a la carpeta de documentación de cada repositorio y se elimina el archivo *".htaccess"* ya que el *rewritebase* por defecto de OnToology tiene que ser modificado para poner la ruta local del servidor. Este archivo permite configurar los contenidos que se van a servir del servidor, tales como RDF, Turtle, etc. Además, se encarga de las redirecciones a las diferentes respuestas http. Con este paso se consigue que no se necesiten realizar modificaciones, salvo la mencionada anteriormente. Una vez realizado esto, la carpeta de

documentación se ubica en el servidor web y comenzaría a estar disponible. Se sugirió como mejora lo siguiente: "Actualmente el proceso se realiza sin automatización, aunque se podría realizar un *webhook* [23] que actualice automáticamente la documentación de OnToology y sobrescriba el archivo mencionado". Un *webhook* permite subscribir por ejemplo, un repositorio a ciertos eventos. Cuando uno de estos eventos ocurre, se manda una petición a la url del webhook. Esto puede usarse para que cada vez que haya cambios en un repositorio se mande una petición de actualización, backup, despliegue en un servidor, etc.

Como última medida, se procedió a generar una *release* para cada uno de los repositorios que albergaban vocabularios, con un número de versión acorde con su estado. Para ello, GitHub nos proporciona herramientas integradas dentro de su plataforma con dicho fin, pudiendo generar estas *releases*[24] de una manera muy sencilla. Esta release será la que se incluya como referencia a la versión más actualizada dentro de la documentación proporcionada por la visualización.

Para finalizar, se muestra una tabla a modo de resumen de los problemas encontrados, ya sea en cuanto a contenido, tecnología o desarrollo del proceso. En ella se muestran los problemas que han sido resueltos y aquellos que no, los cuales deberán resolverse en un futuro.

Problema	*Tipo*	*Resuelto*
Actualización de referencias en las ontologías al migrar de repositorio	Contenido	Parcialmente. Los repositorios de sociedad-bienestar-equipamientos-municipales y sector-publico-organismos no han podido ser actualizados
Resolución de las equivalencias entre ontologías de ambos repositorios	Contenido	Parcialmente. El repositorio sociedad-bienestar-equipamientos-municipales no parece ser equiparable a otra ontología similar
Creación de repositorios no existentes	Desarrollo	Si
Omisión del fichero de ontología en algunos repositorios	Contenido	Si

Actualización del repositorio sector-publico-organismos	Contenido	No
Actualización del repositorio sociedad-bienestar-equipamientos-municipales	Contenido	No
Actualización de vocabularios a publicar	Contenido	Si
Generación de documentación mediante OnToology	Tecnología	Si
Repositorio urbanismo-infraestructuras-callejero genera problemas con su ontología – OOPS!	Tecnología	Si
Generación de nuevos documentos READ.ME que muestren de una manera eficaz las actualizaciones y cambios recientes	Contenido	Si
Regeneración de la documentación publicada	Contenido	Si
Generación de una visualización para los vocabularios	Tecnología	Si
Generación de documentación en multi idioma	Tecnología	No
Generación de documentación a partir de repositorios agregados	Tecnología	Si*
Actualización de documentación de los repositorios a publicar	Contenido	Si
Configuración de propiedades para la generación automática de documentación	Tecnología	Si
Publicación de vocabularios	Desarrollo	Si**
Unificación de formatos en la visualización	Tecnología	No
Enlaces a los repositorios en la visualización	Contenido	No

Adición de descripción a la visualización	Contenido	Si
Generación de releases de los repositorios	Desarrollo	Si

Tabla 8. Tabla resumen de los problemas encontrados y su resolución

** Gracias a Ahmad Alobaid*

*** Con ayuda de Raúl Alcazar*

Capítulo 4

Resultados y conclusiones

Este trabajo ha querido ser origen de la creación y configuración de herramientas y servicios necesarios para la prestación de soporte tecnológico en la creación y publicación de vocabularios de datos abiertos. A lo largo de todo el proceso, han ido surgiendo problemas e inconvenientes; algunos de ellos han sido de fácil solución y otros en cambio, todavía no han podido ser resueltos. Aún así, si se tiene en cuenta la tabla 8, la cual resume en una lista los principales encontrados, el número de problemas que se han resuelto refleja unos buenos resultados al término del análisis.

Cabe destacar la creación y configuración de los repositorios necesarios como soporte para la creación de vocabularios. Durante este proceso se ha observado que una configuración correcta de las herramientas utilizadas permite disponer de documentación actualizada y control de versiones a lo largo de todo el proceso de creación, manteniendo en un mismo espacio todos los archivos necesarios. Se ha visto cómo estas herramientas, OnToology principalmente, nos han provisto de los servicios necesarios para la prestación del soporte tecnológico requerido a lo largo del proceso.

En cuanto al proceso de publicación, se destaca la obtención y regeneración satisfactoria de la documentación publicada anteriormente como primer paso para la configuración del servicio necesario para su visualización. En esta última parte del trabajo, se ha visto como se pueden crear y configurar repositorios agregadores de tal forma que herramientas como OnToology nos proporcione este servicio de visualización. Por último, se han generado *releases* para cada vocabulario; lo que sin duda ayudará a la reutilización y desarrollo de los mismos.

Teniendo en cuenta los resultados obtenidos, pueden extraerse una serie de conclusiones. En primer lugar, todo el proceso de creación y puesta en marcha de conjuntos de datos abiertos para su posterior publicación, es un proceso muy complejo que tiene la necesidad de realizar un análisis a priori de los conjuntos de datos en

profundidad, las herramientas y plataformas de que se disponen y una serie de estándares y buenas prácticas. En segundo lugar, se ha llegado a la conclusión de que es de vital importancia mantener la coherencia a la hora de realizar un proceso de creación y desarrollo, realizando una actualización constante de la documentación y llevando un control de versiones eficaz; manteniendo en todo momento una estructura y organización coherente de los archivos del proyecto. Por último, se ha conseguido realizar un buen trabajo en función de los objetivos marcados, aunque siempre hay cabida para futuras mejoras.

Capítulo 5

Líneas futuras

A pesar del esfuerzo dedicado a la realización del presente trabajado y debido a las limitaciones de tiempo disponible, hay problemas que aún necesitan ser resueltos. Además, hay una serie de tareas deseables que lograrían mejorar aun más el proyecto realizado.

En primer lugar, se deberían resolver cuanto antes los problemas no resueltos. Algunos de los vocabularios que se mantienen en los repositorios no han podido ser actualizados correctamente, debido a cambios en la especificación de ciertas URIs utilizadas. En otros en cambio, no se ha podido establecer una equivalencia entre varias versiones del mismo vocabulario.

En segundo lugar, una descripción de futuras tareas deseables que se podrían llevar a cabo es la siguiente:

- Multi-idioma: Es deseable que en un futuro, la documentación generada por OnToology pudiera mostrarse en diferentes idiomas, según el usuario. Por defecto, esta se genera en inglés, por lo que en un primer paso podría implementarse el español y posteriormente ir incluyendo nuevos idiomas.

- Unificación de formatos: Sería deseable que la página de visualización de los vocabularios mantuviera una sola copia de cada vocabulario con los múltiples formatos en los que se encuentra disponible. Actualmente se muestra una copia de cada vocabulario por cada uno de los formatos.

- Enlazado: Es deseable que en un futuro, la pagina de visualización de los vocabularios permitiera un nuevo campo para cada vocabulario. Este nuevo

campo contendría un enlace al repositorio donde se encuentra ubicado dicho vocabulario.

- Rapidez en la generación: Se podría realizar mejoras en el apartado de los repositorios agregados expuesto. Para ello, si los repositorios que forman parte del repositorio agregador ya disponen de documentación, no generar de nuevo la documentación del repositorio agregador sino solo reutilizar dicha documentación; lo que conlleva más rapidez en la generación.

- Servidor de integración continua: Se podría estudiar la posibilidad de realizar un enlazado de los repositorios con algún servidor de integración continua tipo TRAVIS o Semaphore para generar información cada vez que haya una *release*.

- Automatización del proceso de publicación: Cabe la posibilidad de estudiar el proceso de publicación de vocabularios para descubrir nuevas formas de publicar de una manera más sencilla los vocabularios con todas sus redirecciones correspondientes.

Bibliografía

[1] - Título IV. Del Gobierno y de la Administración - Constitución Española. (2018). Congreso.es. From http://www.congreso.es/consti/constitucion/indice/titulos/articulos.jsp?ini=105&tipo=2

[2] - Acerca de la iniciativa Aporta | datos.gob.es. (2018). Datos.gob.es. From http://datos.gob.es/es/acerca-de-la-iniciativa-aporta

[3] - FEMP - Federación Española de Municipios y Provincias", Femp.es, 2018. [Online]. Available: http://www.femp.es/.

[4] - DATOS ABIERTOS Guía estratégica para su puesta en marcha Conjuntos de datos mínimos a publicar, 8th ed. GRUPO DE DATOS ABIERTOS DE LA RED DE ENTIDADES LOCALES POR LA TRANSPARENCIA Y PARTICIPACIÓN CIUDADANA 1, 2017.

[5] - What is a Vocabulary? W3.org. (2018). Ontologies - W3C. [online] Available at: https://www.w3.org/standards/semanticweb/ontology

[6] - Research, S. (2018). protégé. [online] Protege.stanford.edu. Available at: https://protege.stanford.edu/

[7] - RDF - Semantic Web Standards, W3.org, 2018. [Online]. Available: https://www.w3.org/RDF/

[8] - W3.org. (2018). RDF Schema 1.1. [online] Available at: https://www.w3.org/TR/rdf-schema/

[9] - Vista General del Lenguaje de Ontologías Web (OWL), W3.org, 2018. [Online]. Available: https://www.w3.org/2007/09/OWL-Overview-es.html

[10] - W3.org. (2018). Describing Linked Datasets with the VoID Vocabulary. [online] Available at: https://www.w3.org/TR/void/#dataset

[11] - W3.org. (2018). Data - W3C. [online] Available at: https://www.w3.org/standards/semanticweb/data

[12] - Group, O. (2018). The Open Definition - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge. [online] Opendefinition.org. Available at: https://opendefinition.org/

[13] - Build software better, together, GitHub, 2018. [Online]. Available: https://github.com/about

[14] - OnToology, Ontoology.linkeddata.es, 2018. [Online]. Available: http://ontoology.linkeddata.es/about

[15] - Red Temática - Bienvenido, Opencitydata.es, 2018. [Online]. Available: http://opencitydata.es/web/guest

[16] - AENOR - Asociación Española de Normalización y Certificación, Aenor.es, 2018. [Online]. Available: http://www.aenor.es/aenor/inicio/home/home.asp

[17] - AENOR: Norma UNE 178301:2015, Aenor.es, 2018. [Online]. Available: http://www.aenor.es/aenor/normas/normas/fichanorma.asp?tipo=N&codigo=N0054318&PDF=Si#.WxJpEUiFOUk

[18] - #OJOALDATA100 - Medialab-Prado Madrid, Old.medialab-prado.es, 2018. [Online]. Available: http://old.medialab-prado.es/article/ojoaldata100

[19] - R. Stevens and U. Sattler, "Disjointness Between Classes in an Ontology", Ontogenesis, 2018. [Online]. Available: http://ontogenesis.knowledgeblog.org/1260.

[20] - OOPS! - OntOlogy Pitfall Scanner!, Oops.linkeddata.es, 2018. [Online]. Available: http://oops.linkeddata.es/.

[21] - Git - Submódulos, Git-scm.com, 2018. [Online]. Available: https://git-scm.com/book/es/v1/Las-herramientas-de-Git-Subm%C3%B3dulos.

[22] - The most popular self-hosted file share and collaboration platform, Nextcloud, 2018. [Online]. Available: https://nextcloud.com/

[23] - Webhooks | GitHub Developer Guide, Developer.github.com, 2018. [Online]. Available: https://developer.github.com/webhooks/

[24] - Creating Releases - User Documentation, Help.github.com, 2018. [Online]. Available: https://help.github.com/articles/creating-releases/

Este documento esta firmado por

Firmante	CN=tfgm.fi.upm.es, OU=CCFI, O=Facultad de Informatica - UPM, C=ES
Fecha/Hora	Wed Jun 06 09:49:02 CEST 2018
Emisor del Certificado	EMAILADDRESS=camanager@fi.upm.es, CN=CA Facultad de Informatica, O=Facultad de Informatica - UPM, C=ES
Numero de Serie	630
Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)