Wikidata y *DBpedia*: viaje al centro de la web de datos

Wikidata and *DBpedia*: A journey to the center of a web of data

Tomás Saorín y Juan-Antonio Pastor-Sánchez

Saorín, Tomás; **Pastor-Sánchez, Juan-Antonio** (2018). "*Wikidata* y *DBpedia*: viaje al centro de la web de datos". *Anuario ThinkEPI*, v. 12, pp. 207-214.

https://doi.org/10.3145/thinkepi.2018.31

Publicado en *IweTel* el 14 de diciembre de 2017

Resumen: A partir de *Wikipedia*, como fuente de conocimiento organizado en forma de artículos enciclopédicos, editada mediante la colaboración masiva online, se han desarrollado dos proyectos de carácter semántico: *DBpedia* y *Wikidata*. Se analizan las diferencias y similitudes entre ambos modelos de datos y modelo de producción, y se especula sobre la posible evolución y coexistencia de ambos a partir de sus puntos fuertes. Su fortaleza como grafo abierto de conocimiento multidominio aporta un gran valor a la extensión de la web de datos, al actuar como punto de interconexión entre diferentes dominios.

Palabras clave: Web semántica; Ontologías; *Wikipedia*; *Wikidata*; *DBpedia*; Web de datos; Grafo de conocimiento abierto.

Abstract: *DBpedia* y *Wikidata* are two semantic projects built on the top of *Wikipedia* encyclopedic content that has been crowdsourced, created, and maintained by thriving online communities. This article outlines differences and similarities between these two projects, pointing out their data models, curation patterns, and technologies. In addition, it forecasts some possible trends for their co-evolution as cross-domain open knowledge graphs and their significant value as hubs of interlinking datasets of various domains.

Keywords: *Wikidata*; *DBPedia*; Semantic web; Ontologies; Open Knowledge graph.

Introducción

Wikipedia se ha convertido en una fuente de información de referencia utilizada por millones de usuarios sobre conceptos, hechos, ciencia y cultura. Sus contenidos tienen un posicionamiento óptimo, son reutilizados por numerosas aplicaciones y servicios de terceros y conforman el núcleo a partir del cual se alimenta el nodo central de la web de datos enlazados: *DBpedia*.

DBpedia comenzó su andadura en 2007 como una base de conocimiento multidominio, cuya extensión y cobertura fue ampliándose en diversos ciclos (**Bizer** *et al.*, 2009). Es el mismo año del lanzamiento del *linking open data community project* en el *W3C* y el comienzo de la materialización a gran escala de las ideas subyacentes en el término "web semántica", que alcanzó alrededor de los años 2010-2012 un punto cercano a la madurez como tecnologías, prácticas, contenidos y mercados (**Saorín**; **Peset**; **Ferrer-Sapena**, 2013).

En 2012 la *Fundación Wikimedia* presentó *Wikidata* como una base de conocimiento estructurado mantenido de forma colaborativa. Su objetivo es crear una fuente común de datos para su reutilización en otros proyectos *Wikimedia* o por parte de terceros gracias a su licencia *Creative Commons* de dominio público.

Tanto *DBpedia* como *Wikidata* persiguen objetivos similares, pivotan alrededor de un mismo proyecto, *Wikipedia*, pero adoptan procedimientos diferenciados.
http://dbpedia.org
http://www.wikidata.org

Entre ambos proyectos existen una serie de conexiones y objetivos compartidos desde el punto de vista de la organización del conocimiento humano. Analizaremos en esta nota su complementariedad y sus perspectivas de futuro. La disponibilidad de conjuntos de datos estructurados que formalizan y estructuran dicho conocimiento supone una herramienta de incalculable valor. El hecho de que existan herramientas disponibles para su acceso, consulta y reutilización supone un valor añadido, puesto que suma un nivel de abstracción respecto al consumo y reutilización de estos datos.

Formalizando el mundo a través de *Wikipedia*: la ontología *DBpedia*

El proyecto *DBpedia* surge sobre la base de la formalización del conocimiento de los artículos de *Wikipedia*. La formalización es un procedimiento imprescindible para estructurar y representar la semántica de la información en entornos digitales para su procesamiento por máquina.

"Una *infobox* es una plantilla *wiki* en la que se define una estructura de datos común y su representación visual para determinados tipos de artículos (personas, ciudades, películas, etc.)"

Los artículos de *Wikipedia* no sólo incluyen contenido textual; buena parte de ellos también contiene una gran cantidad de información estructurada mediante fichas descriptivas (*infoboxes*). Una *infobox* es una plantilla *wiki* en la que se define una estructura de datos común y su representación visual para determinados tipos de artículos (personas, ciudades, películas, etc.). Tienen un bajo nivel de formalización y carecen de mecanismos automáticos que garanticen su correcto diseño y aplicación o la integridad de los datos que contienen. Tampoco existe una interrelación entre elementos descriptivos equivalentes de distintas *infoboxes*, ni entre *infoboxes* de las ediciones de *Wikipedia* en diferentes idiomas. Todo ello se debe a que las plantillas descriptivas se elaboran como contenidos *wiki*: sobre la marcha, editables, no centralizados.

Desde esta situación surge *DBpedia*, que busca convertir el conocimiento blando de las *infoboxes* de *Wikipedia* en conocimiento formalizado mediante una ontología aplicando los principios y tecnologías de *linked open data* del *W3C*. La creación de la ontología *DBpedia* es un proceso intelectual, desarrollado y mantenido por una comunidad *crowdsourced*, que no se encuentra relacionada directamente con las de *Wikipedia*. Dicha ontología se organiza en una taxonomía de 685 clases y 2.795 propiedades cuyo dominio tiene un alcance universal (*cross-domain*). Lo anterior conlleva que es posible representar ámbitos del conocimiento tan dispares como especies botánicas, ciudades, concursos de televisión, deportistas o personajes de ficción.

"La ontología *DBpedia* se organiza en una taxonomía de 685 clases y 2.795 propiedades cuyo dominio tiene un alcance universal (*cross-domain*)"

La ontología *DBpedia* se integra en el ecosistema de vocabularios enlazados (*linked open vocabularies*) de dos formas:
- reutilizando elementos de ontologías y esquemas de metadatos ya existentes (SKOS, DCTERMS, FOAF, etc.);
- definiendo equivalencias con clases y propiedades de otros vocabularios.

Por ejemplo, la clase "Person" está definida en *DBpedia* como equivalente a "Person" en *Schema.org*[1].

Poblar el mundo: ¿de dónde salen los datos de DBpedia?

Aunque la ontología de *DBpedia* es una sola, los datos provienen de las ediciones de *Wikipedia* de diferentes idiomas. Las clases de *DBpedia* se han mapeado con *infoboxes* concretas de *Wikipedia*, contemplando la realidad multilingüe de *Wikipedia*[2]. Un ejemplo sería la clase de *DBpedia* "Road" que se encuentra mapeada, entre otras, con las plantillas "Infobox road" (edición en inglés) y "Ficha de calle" (edición en español). De este modo se identifican las clases de *DBpedia* sobre las que pueden aplicarse los datos estructurados de cada *infobox* de *Wikipedia*.

La heterogeneidad respecto a la granularidad de las *infoboxes* de las diferentes ediciones de *Wikipedia* supone un problema añadido. Algunas ediciones de *Wikipedia* tienen plantillas muy específicas que se corresponden con clases igual de específicas en *DBpedia*. Sin embargo, otras ediciones carecen de tal especificidad y se utilizan plantillas muy generales. Un ejemplo de ello sería

el artículo sobre el ajedrecista cubano José Raúl Capablanca: mientras que en la edición inglesa se utiliza la plantilla "Infobox chess player" en la española se usa "Ficha de persona", al tiempo que en la ontología *DBpedia* está disponible la clase "chessPlayer".

Algo más complejo es el proceso para definir equivalencias entre las propiedades de la ontología y los campos de las plantillas debido al carácter semi-estructurado y altamente personalizable de estas últimas. No todos los campos de las plantillas pueden mapearse contra alguna de las propiedades de *DBpedia*. Muchos de los campos de las plantillas son excesivamente específicos e incluso existen ciertas redundancias puesto que al fin y al cabo las *infoboxes* están ideadas para facilitar la consulta de los artículos de *Wikipedia* por parte de los lectores, más que como una herramienta de formalización.

El mapeado de *DBpedia* se aplica a través de una herramienta de extracción automática de información a partir del contenido de *Wikipedia*[3]. De este modo es posible obtener grafos RDF de cada artículo de *Wikipedia* y su posterior integración en el conjunto de datos de *DBpedia*. Sin duda es una tarea con algunas limitaciones, en donde la supervisión y revisión por parte de la comunidad *DBpedia* resulta imprescindible, pero que permite procesar una gran cantidad de contenidos para generar un gran volumen de datos estructurados.

El *dataset* de *DBpedia* tiene un alto grado de integración en el ecosistema *linked open data*. En este sentido *DBpedia* no sólo conserva los vínculos entre diferentes recursos del conjunto de datos (incluyendo tanto a los recursos como las categorías) sino que también incluye enlaces a recursos externos tales como páginas web o clases de otras ontologías como *Yago*. Sin embargo, el *dataset* de *DBpedia* no incluye los enlaces a las autoridades de los artículos de *Wikipedia* (VIAF, GND, BNF, *Worldcat*, *Library of Congress Control Number*, etc.). Lo anterior limita en cierto modo las posibilidades de conexión con otros datasets, generalmente bibliotecarios, que vinculan sus registros con los vocabularios de control de autoridades.

¿Y si lo hacemos entre todos? *Wikidata*: la central de datos para *Wikipedia*

DBpedia puso de relieve el potencial del conocimiento formalizado e interrogable de *Wikipedia*. No era raro que la propia comunidad *Wikimedia* se preguntara si no era mejor producir datos en origen en lugar de obtenerlos a través de una extracción forzada. La respuesta a este impulso semántico fue *Wikidata*, puesta en marcha como infraestructura compartida por todas las ediciones de *Wikipedia* para datos factuales. Es un proyecto análogo a lo que ha supuesto el *Commons* para las imágenes y medios.

Wikidata es un *hub* de datos estructurados en el que cada entidad se representa mediante una IRI desreferenciable donde también se recogen los enlaces a los artículos equivalentes de *Wikipedia* en diferentes idiomas: la entidad *Q770676* de *Wikidata* representa el hecho "Spain 12–1 Malta" -el histórico partido de clasificación para la Eurocopa entre las selecciones de fútbol de España y Malta en 1983- del que existen artículos en las ediciones en español, catalán, inglés, turco y japonés)[4].

¿Cómo se vinculan los artículos con *Wikidata*? Al crear un artículo enciclopédico también debe crearse el item correspondiente en *Wikidata*, donde se define al menos una etiqueta en un idioma, una mínima descripción para desambiguar y un *sitelink* al correspondiente artículo o artículos en cada edición de *Wikipedia*.

En *Wikidata* la descripción de las entidades implican el desarrollo de un proceso de formalización colaborativa del conocimiento (**Vrandečić**; **Krötzsch**, 2014). En *Wikidata* no sólo se definen los datos, sino también las propiedades para almacenarlos. Hay más de 4.700 propiedades agrupadas en 8 ámbitos:

- Generic
- Person
- Organization
- Events
- Works
- Terms
- Geographical feature
- Others

El momento clave para *Wikidata* es el paso a su uso para construir las *Infoboxes* en *Wikipedia*

Hasta ahora los datos de las *infoboxes* de *Wikipedia* y los datos de *Wikidata* están separados y sin sincronización. Un primer paso intermedio es el que actualmente encontramos en cada *infobox*: la indicación de "[editar datos en *Wikidata*]" al final de cada ficha permite a los editores voluntarios corregir y completar datos en *Wikidata* pero no en el artículo de *Wikipedia* donde deben volver a introducirse en la *infobox*.

Ya se está probando experimentalmente la incorporación automática de los datos de *Wikidata* en las *infoboxes*. De este modo los datos ya no estarán en el texto del artículo, sino que serán una referencia a *Wikidata*, completando de esta forma el círculo para el que se diseñó la plataforma.

La diferencia puede verse claramente en los

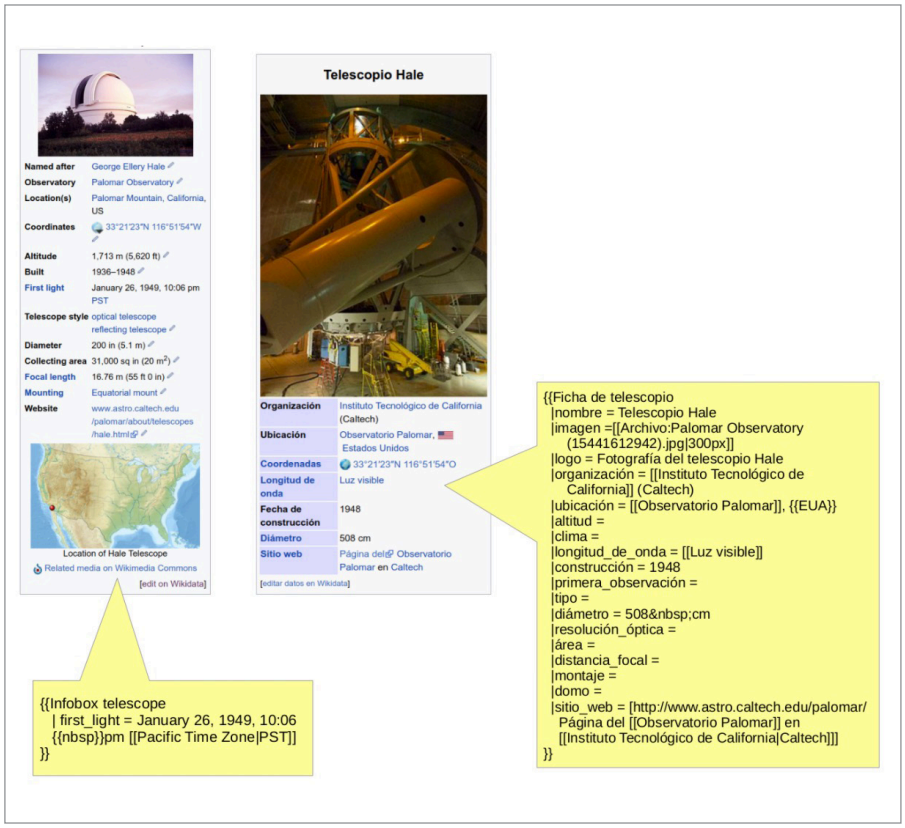

Figura 1. Inclusión de la plantilla de telescopios en la edición inglesa mediante *Wikidata* (izquierda) y en la edición en español con la plantilla clásica (derecha)

artículos sobre telescopios de la edición inglesa. La generación de cada *infobox* a partir de los datos extraídos de *Wikidata* se realiza insertando en el artículo el código de plantilla {{infobox telescope}}[5]. En la definición de la plantilla se invocan las propiedades de *Wikidata* necesarias. Un ejemplo de uso es el artículo sobre el "Hale Telescope". Sin embargo, el artículo equivalente en la edición española "Telescopio Hale" sigue utilizando una plantilla clásica en la que el editor, junto a cada campo, tiene que indicar expresamente sus valores, como contenido *wiki*, y no datos procedentes de *Wikidata* (figura 1). Este enfoque es compatible con el hecho de que un editor pueda incorporar información local cuando el dato no está en *Wikidata* o desea sobreescribirlo. Si no se indica lo contrario, la plantilla toma los datos de las propiedades declaradas en *Wikidata*.

Wikidata también tiene un gran potencial en su *Query Service*[6], un *Sparql Endpoint* donde además de recuperar datos, estos pueden utilizarse para representar mapas, diagramas, líneas de tiempo, listados y otras visualizaciones convenientes. Sin duda, cuando estas visualizaciones dinámicas se integren en los artículos de *Wikipedia*, se producirá un gran salto en la presentación de los datos incrementando la expresividad del contenido de los artículos.

Otra de las posibilidades es la capacidad de *Wikidata* para generar listados automáticos en *Wikipedia*. Los listados son elementos frecuentes, y tanto su actualización como su edición en *wikitexto* presentan complicaciones. Herramientas como *Listeria!*, permiten realizar esta tarea de forma consistente en varias *Wikipedias* al mismo tiempo. Basta con insertar la plantilla {{Wikidata list}} en un artículo, y un *bot* consultará los datos a través del *Query Service* de *Wikidata* y mantendrá el listado actualizado en el artículo. De esta forma los datos en *Wikidata* se propagan a las *Wikipedias*[7].

En los artículos de personas de *Wikipedia* encontramos al final una ficha de Control de autoridades, cuya plantilla toma los datos registrados en el elemento de *Wikidata* correspondiente[8] y ya no requiere la introducción de ningún identificador en los artículos. Se trata de un subconjunto de propiedades en *Wikidata* denominadas *Identifiers* (datos del tipo "Identificadores externos", dentro de la clase "*Wikidata property for authority control*"). Esto permite enlazar las entidades con los principales registros de autoridad en un enfoque muy propio de *linked open data* (**Pastor-Sánchez**; **Saorín**, 2015).

¿Hay diferencias entonces entre *DBpedia* y *Wikidata*?

A estas alturas el lector puede pensar que ambos proyectos son esencialmente lo mismo, que se producen usando diferentes técnicas y procesos, pero que sus resultados son los mismos, datos estructurados derivados de los artículos de *Wikipedia*.

Para evitar llegar a esta conclusión, se expone el esquema comparativo del trabajo "Wikidata through the eyes of *DBpedia*" (**Ismayilov** *et al.*, 2016)

Para comprender mejor las implicaciones de la tabla 1 conviene precisar que el modelo de datos abstracto de *Wikidata* es complejo e incluye dos tipos de entidades: **ítems y propiedades**. En

Tabla 1. Diferencias entre *DBpedia* y *Wikidata*

Aspecto	*DBpedia*	*Wikidata*
Identificadores (URI, IRI)	Usa identificadores legibles por humanos a partir del título de los artículos en cada idioma.	Usa identificadores numéricos independientes del idioma
Estructura	Usa RDF de forma nativa en su modelo datos.	Desarrolla su propio modelo de datos, con mayor capacidad para representar la procedencia de los datos (*provenance*). Permite generar diferentes serizaliaciones en RDF
Esquema	Ambos esquemas están construidos por una comunidad y son multilingües. La ontología de *DBpedia* se basa en OWL para organizar los datos que extrae e integra de las diferentes ediciones de *Wikipedia*.	El esquema de *Wikidata* evita el uso directo de términos de RDFS y OWL y redefine la mayoría de ellos (Por ejemplo *lawkd:P31* define una propiedad local que es similar a *rdf:type*. Existen, no obstante, intentos de conectar las propiedades de *Wikidata* con RDFS/OWL.
Curación	Los datos de *DBpedia* se extraen por procesos automáticos de *Wikipedia* y constituyen un *dataset* de solo lectura. Los editores de *Wikipedia* son, indirectamente, los curadores de los datos de *DBpedia*. Por su naturaleza semiestructurada en *wiki*, no pueden capturarse todos los datos y pueden producirse errores durante la extracción.	*Wikidata* tiene su propio entorno de edición de contenidos, *WikiBase*, que permite crear, editar y depurar tanto sus datos como su estructura.
Publicación	Ambos *datasets* se publican mediante técnicas *linked data*, incluyendo *datasets dumps*, URIs derreferenciables y *Sparql endpoints*.	*Wikidata* además se ofrece con un potente entorno de visualización de resultados, que permite obtener mapas, diagramas o gráficos a partir de los datos.
Cobertura	*DBpedia* proporciona identificadores para todos los elementos estructurales de una edición de *Wikipedia*. Incluye artículos, categorías, redirecciones y plantillas.	*Wikidata* crea identificadores comunes para conceptos que existen en más de un idioma. No todos los artículos, categorías, plantillas y *redirects* de una edición tiene un elemento *Wikidata*. *Wikidata* permite crear items para conceptos que no se corresponden con elementos de los proyectos *Wikimedia* (por ejemplo, fuentes bibliográficas)
Actualización de los datos	*DBpedia* es un *dataset* estático, de sólo lectura y de actualización periódica aproximadamente semestral. Como excepción existe *DBpedia Live*, basado en el proceso de copias locales de ediciones de *Wikipedia* (en inglés, francés y alemán).	*Wikidata* es editable, y los editores pueden crear, actualizar y corregir datos sobre la marcha.

ambos casos tienen asignados identificadores IRI únicos[9] de forma que tanto la definición de los datos en sí, como de las propiedades que los describen se encuentran autocontenidas en *Wikidata*. Para cada propiedad de una entidad, pueden predicarse diferentes valores, indicar su prioridad, añadir cualificadores, incluir aproximaciones y consignar la fuente de procedencia del dato. A un nivel más básico, el modelo de datos de *Wikidata* concuerda con RDF: *type*, *subclassof*, *subpropertyof*, *object*, *subject*, etc.[10]

Wikidata proporciona una infraestructura local para mejorar los otros proyectos *Wikimedia*. Su publicación con licencia CC0 (dedicación al dominio público) y su forma de publicación como *linked open data* (*Sparql*, DUMPs y APIs) permite su uso por terceros, multiplicando el valor primario de los proyectos *Wikimedia* a través de la reutilización masiva de datos abiertos. Por su parte, *DBpedia* es, por diseño, una infraestructura abstracta de datos abiertos, orientada primariamente a la reutilización y consumo de datos mediante una ontología.

Conclusiones y prospectiva: dos anillos para dominarlos a todos

En la web semántica no hay un solo actor ni un solo centro. Para *Wikidata* se ha usado en tono admirativo el calificativo de "nueva piedra Rosetta", pero también la analogía tolkiana de "un anillo para dominarlos a todos" que nos recuerda lo peligroso de un solo poder dominante (**Kolbe**, 2015; **Hinojo**, 2015). ¿Pueden convivir estos dos proyectos o el mundo es como ese pueblo del *Far West* que era demasiado pequeño para dos pistoleros?

Si el principio de la web semántica era la asunción de que el mundo es abierto y que cualquiera puede decir cosas sobre cualquier tema -resumido en el slogan AAA, "*Anyone can say anything about any topic*" (**Allemang**; **Hendler**, 2011)- en el caso de *DBpedia* y *Wikidata* sus afirmaciones son las mismas, puesto que los datos son los mismos. ¿Qué cambia? Podríamos hablar de añadir otra A al slogan, de forma que también se puede decir "de cualquier forma" *-Anyway-* haciendo referencia a diferentes formalizaciones, las cuales también puede ser reconciliadas entre sí

usando mecanismos disponibles en las tecnologías de la web semántica.

Wikidata es una capa transparente, pero imprescindible para la viabilidad de *Wikipedia* a largo plazo y para su infraestructura de conocimiento estructurado. No obstante, la incorporación de datos de *Wikidata* en los artículos de *Wikipedia* requiere que la primera alcance un nivel de madurez homogéneo.

"La publicación de *Wikidata* con licencia CC0 y su forma de publicación como *linked open data* permite su uso por terceros, multiplicando el valor primario de los proyectos *Wikimedia* a través de la reutilización masiva de datos abiertos"

Posiblemente nos encontramos ante el comienzo de un camino en el que *Wikidata* y *DBpedia* se aproximen. Tal vez con el tiempo la primera se convierta en una fuente de datos para la segunda, superando la extracción de las *infoboxes* de *Wikipedia*. En ese momento *DBpedia* se convertirá en una metaestructura para la organización y el consumo operativo de los datos de *Wikidata*. La vinculación que puede encontrarse en *DBpedia* entre los recursos de esta y las entidades de *Wikidata* ya es un primer escalón en este proceso. En cualquier caso, ambos proyectos son necesarios y complementarios puesto que está demostrándose que conforman el núcleo de la esfera de la web de datos.

Notas

1. Ver en
http://dbpedia.org/ontology/Person

la definición de la equivalencia con
http://schema.org/Person

2. Se pueden consultar dichos mapeados en:
http://mappings.dbpedia.org

3. Existe una amplia documentación sobre *DBpedia Information Extraction Framework* en:
http://wiki.dbpedia.org/documentation

4. En *Wikidata* esta entidad se identifica mediante la IRI:
http://www.wikidata.org/entity/Q770676

5. Ver ejemplo en:
https://en.wikipedia.org/wiki/Template:Infobox_telescope

6. Este servicio está disponible en
https://query.wikidata.org

7. Se trata de una herramienta experimental operativa en las ediciones inglesas y alemana. El *ListeriaBot* está disponible en
https://tools.wmflabs.org/listeria

8. Ver la plantilla:
https://es.wikipedia.org/wiki/Plantilla:Control_de_autoridades

9. Dos ejemplos de IRI serían:

-para "Spain 12–1 Malta":
http://www.wikidata.org/entity/Q770676

-para la propiedad "location":
http://www.wikidata.org/entity/P276

10. A este respecto consultar:
https://www.wikidata.org/wiki/Wikidata:Relation_between_properties_in_RDF_and_in_Wikidata

Referencias

Allemang, Dean; **Hendler, Jim** (2011). *Semantic web for the working ontologist: Effective modeling in RDFS and OWL*. San Francisco, CA: Morgan Kaufmann; Oxford: Elsevier Science. ISBN: 978 0123859655

Bizer, Christian; **Lehmann, Jens**; **Kobilarov, Georgi**; **Auer, Sören**; **Becker, Christian**; **Cyganiak, Richard**; **Hellmann, Sebastian** (2009). "DBpedia - A crystallization point for the web of data". Web semantics: Science, services and agents on the world wide web, v. 7, n. 3, pp. 154–165.
http://www.websemanticsjournal.org/index.php/ps/article/view/164/162
https://doi.org/10.1016/j.websem.2009.07.002

Hinojo, Àlex. (2015). "*Wikidata*, la nova pedra de Rosetta". *CCCBLAB. Investigació i innovació en cultura*, 25 nov.
http://lab.cccb.org/ca/la-nova-pedra-de-rosetta

Ismayilov, Ali; **Kontokostas, Dimitris**; **Auer, Sören**; **Lehmann, Jens**; **Hellmann, Sebastian** (2016). **"**Wikidata through the eyes of DBpedia". *Semantic web*, v. 0, n. 0, pp. 1–11.
http://www.semantic-web-journal.net/content/wikidata-through-eyes-dbpedia-1

Kolbe, Andreas (2015). "Whither Wikidata?". *Wikipedia singpost*, 2 December.
https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2015-12-02/Op-ed

Pastor-Sánchez, Juan-Antonio; **Saorín, Tomás** (2015). "Web semántica. Informe de situación 2014". *Informes ThinkEPI 2015 sobre documentación y comunicación*, v. 1, pp. 177-188.
http://dx.doi.org/10.3145/info.2015.12

Saorín, Tomás; **Peset, Fernanda**; **Ferrer-Sapena, Antonia** (2013). "Factores para la adopción de *linked data* e implantación de la web semántica en bibliotecas, archivos y museos". *Information research*, v. 18, n. 1, paper 570.
http://InformationR.net/ir/18-1/paper570.html

Vrandečić, Denny; **Krötzsch, Markus** (2014).

"Wikidata: A free collaborative knowledge base". *Communications of the ACM*, v. 57, n. 10, pp. 78-85.
https://static.googleusercontent.com/media/research.google.com/es//pubs/archive/42240.pdf
http://dx.doi.org/10.1145/2629489

Tomás Saorín
Universidad de Murcia
Facultad de Comunicación y Documentación
Departamento de Información y Documentación
tsp@um.es

Juan-Antonio Pastor-Sánchez
Universidad de Murcia
Facultad de Comunicación y Documentación
Departamento de Información y Documentación
http://webs.um.es/pastor
pastor@um.es

* * *

Un ejemplo de datos inconsistentes en Wikipedia

Josu Aramberri

Un caso de estudio actual, que podéis añadir:
En la *Wikipedia* en español, el piloto Aleix Espargaró figura con el dato de "nacionalidad" Español:
https://es.wikipedia.org/wiki/Aleix_Espargaró

En la *Vikipèdia* en catalán, la "nacionalitat" del mismo piloto es Catalunya:
https://ca.wikipedia.org/wiki/Aleix_Espargaró_i_Villà

Supongo que *Wikidata*, al recoger "datos factuales", eliminaría estas inconsistencias.

jaramberri@arija.org

* * *

Wikidata como hub de datos para *Wikipedia*

Tomàs Saorín

Precisamente tu observación apunta a una de las peculiaridades del diseño de *Wikidata* como *hub* de datos para *Wikipedia*. El diseño teniendo en cuenta el modelo de trabajo *wiki* implica que la decisión sobre qué datos son los adecuados para un artículo sea de los editores, de la comunidad, mediante consenso. La existencia de *Wikidata* como banco de datos central choca con la capacidad que debería tener la comunidad editora catalana para asignar las nacionalidades mediante un proceso de búsqueda del punto de vista neutral, fuentes verificables, consenso…

Se plantean aquí dos situaciones:

- mantener la autonomía de cada edición de *Wikipedia* para decidir sobre sus contenidos (ámbito *Wikipedia*);
- gestionar la disparidad de criterio a la hora de asignar valores a los datos factuales, por muy objetivos que puedan parecer (ámbito *Wikidata*).

¿Qué nacionalidad aparecería en *Wikidata*? Para los datos de *Wikidata* se exigen criterios parecidos a los de *Wikipedia*: notabilidad, verificabilidad, consenso. Por lo tanto, se permite la discusión sobre el dato, hasta buscar la mayor precisión y el máximo acuerdo.

Pero lo bello es que el modelo de datos admite un buen margen de desacuerdo: la misma propiedad puede declararse con valores distintos si hay disparidad de criterio: por ejemplo una fecha de nacimiento disputada, indicando en cada caso la fuente de autoridad utilizada para afirmarlo. Además, puede matizarse con una aproximación (hacia 1590) para dataciones y esas cosas. Y además a los diferentes valores se les pueden asignar diferentes rankings que establecen el nivel de prioridad con el que serían presentados. Digamos que el 90% de los usuarios optan por la nacionalidad española y un 10% por la catalana. *Wikidata* refleja los dos puntos de vista, pero cualifica la opinión mayoritaria.

Además, las implementaciones que se están haciendo de los *infoboxes* automáticos contra *Wikidata* permiten al editor no estar sometido al dato global: puede optar por usar los datos de *Wikidata*, pero sobreescribir con un valor local cierto campo si lo considera adecuado. Es una situación que origina cierto conflicto, pero que permite mantener el ámbito de decisión dentro del proyecto de la enciclopedia y no en el de los datos.

Por otro lado, las consultas a *Wikidata* también puede solicitar que se obtenga el dato más probable o más aceptado o con mejor ranking, por lo que se ha tratado de implementar la negociación y la flexibilidad en la interpretación de la realidad, aunque sea para datos factuales u objetivos.

tsp@um.es

* * *

¿Se pueden establecer hechos probados por consenso?

Josu Aramberri

No deja de sorprenderme esta interpretación de los "datos factuales" para una enciclopedia como *Wikipedia*.

Entendía hasta ahora que *Wikipedia* refleja hechos probados, no opiniones. No sabía que se podía establecer como hecho probado por "consenso» de una comunidad una característica como la "nacionalidad catalana", cuando no existe ningún organismo oficial ni ningún documento oficial (tarjeta de identidad, pasaporte), que sirva para otorgar la nacionalidad "catalana".

En *Wikidata* la propiedad p27 también está sometida a una interpretación "sui generis" en la definición en catalán.

Me sorprende cómo se pueden crear realidades paralelas para satisfacer los sentimientos y la propaganda, y alterar los datos objetivos, aunque no estén soportados por la razón y los hechos. Así como los nombres pueden tener variantes (e incluso alias), y las fechas pueden ser aproximadas o estimadas, el caso que nos ocupa no está sujeto a principios de incertidumbre como si se tratase de una propiedad termodinámica. Ni debería de cambiar de valor de una *Wikipedia* a otra.

En cualquier caso, gracias por tu detallada explicación.

jaramberri@arija.org

* * *

Modelo de *wikidatos,* enfoques sesgados y resultados razonables

Tomàs Saorín

Creo que los dos estamos de acuerdo en que la entrada sobre la nacionalidad catalana que pones como ejemplo es un disparate. Refleja el momento pasado de rosca que vivimos, y cómo puede afectar esto a un entorno basado en los equilibrios entre puntos de vista como es *Wikipedia*. Nos guste o no, el hecho real es que "la comunidad" (al menos un editor, con el permiso de otros) ha optado por aplicar un criterio de nacionalidad imaginaria y no administrativo. Entendemos que con la masa atómica del sodio no pasa eso, pero sí con otros datos. La existencia de datos objetivos y totalmente factuales no supera ciertas pruebas, muchos datos surgen de un consenso académico y no hay un instituto estadístico que decida. Sí lo hay, pero no en *Wikipedia*, a menos que sea el *Common Sense Institute* con sede en el reino de Redonda.

Lo que yo venía a plantear era que el modelo *Wikidata* aborda de cara estas situaciones, puesto que permite negociar los significados. No sólo en los datos, también la ontología: las propiedades se crean, se debate su alcance y sus contradicciones. Construir una ontología parece fácil, pero implica un modelo formalizado del mundo y eso no es fácil conseguir en modo *crowd*. Que haya mecanismos para la discusión, para el establecimiento de prioridades, para aceptar cierta disidencia, visiones alternativas o visiones disparatadas, permite fluir...

Podríamos quedarnos con que el modelo de *wikidatos* permite que ciertos enfoques muy sesgados entren en el juego, y que el resultado final sea razonable si la propia comunidad alcanza acuerdos razonables. Y una realidad paralela se mantendrá mientras la propia comunidad no sea capaz de revertirla, pero no es para siempre. También podríamos contar con que en el caso de *Wikidata* los debates sobre significados son más globales, y menos afectados por rarezas locales, puesto que deben ser acordadas por una comunidad global. El resto del mundo puede optar por el valor España (*Preferred rank*) y las provincias catalanas opten por pelear también por Catalunya (que terminará siendo un deprecated rank).

A poco que removamos algunas cosas en *Wikipedia*, terminamos en debates sobre la verdad y la objetividad, que no son cosa menor.

tsp@um.es

DIGIBÍS

Digitalización enriquecida

Software de gestión digital
para Archivos, Bibliotecas,
Museos, Exposiciones temporales,
Centros de Documentación

Con
metadatos
ajustados a
la normativa
internacional

Aplicaciones LOD
con Reconciliación
Semántica

Aplicaciones con Recolector
y Repositorio OAI-PMH

Objetos digitales recolectables por
Hispana, Europeana, OAister

No hace falta viajar a la luna para dar a los datos la mayor visibilidad

Un concepto de digitalización y unas aplicaciones que hacen más eficiente el trabajo de las instituciones de memoria

ISO 9001 BUREAU VERITAS Certification
Nº ES042816-1

DIGIBÍS. C/ Alenza, 4. Madrid. Tel.: 914 32 08 88. E-mail: digibis@digibis.com

www.digibis.com

PYME INNOVADORA
Válido hasta el 31 de diciembre de 2018

AENOR
Producto Certificado
Proyecto I+D+i