Wikidata como herramienta para elaborar ontologías y vocabularios controlados

Wikidata as a tool to create ontologies and controlled vocabularies

Juan-Antonio Pastor-Sánchez

Pastor-Sánchez, Juan-Antonio (2021). "*Wikidata* como herramienta para elaborar ontologías y vocabularios controlados". *Anuario ThinkEPI*, v. 15, e15f01.

https://doi.org/10.3145/thinkepi.2021.e15f01

Publicado en *IweTel* el 13 de julio de 2021

Juan-Antonio Pastor-Sánchez
https://orcid.org/0000-0002-1677-1059

Universidad de Murcia
Facultad de Información y Documentación
Depto. de Información y Documentación
Campus Universitario
30100 Espinardo (Murcia), España
pastor@um.es

Resumen: Se muestran las posibilidades que ofrece *Wikidata* para la creación de ontologías y vocabularios controlados sobre dominios de conocimiento específicos. El primer paso consiste en la exploración de las propiedades utilizadas por los ítems de *Wikidata* para definir la pertenencia a un dominio de conocimiento. Posteriormente es posible recuperar la estructura de clases y superclases a las que pertenecen los ítems del dominio. Finalmente se recuperan todas las propiedades de *Wikidata* utilizadas para describir los ítems. El autor incluye una serie de reflexiones sobre el cambio de paradigma en el campo de los vocabularios controlados, puesto que en la actualidad se requiere un alto grado de integración de estos instrumentos en entornos de datos y contenidos digitales. Esta realidad implica una participación de los vocabularios controlados en procesos lógicos de análisis de datos. Por este motivo se requiere un enfoque más dinámico y cercano a las ontologías que a los lenguajes documentales tradicionales.

Palabras clave: *Wikidata*; Ontologías; Vocabularios controlados; Organización del conocimiento.

Abstract: The possibilities that *Wikidata* offers for the creation of ontologies and controlled vocabularies for specific knowledge domains are presented. The first step consists of exploring the properties of *Wikidata* items that are used to define their membership to a knowledge domain. It then becomes possible to retrieve the structure of the classes and superclasses to which the domain items belong. Finally, all the *Wikidata* properties used to describe the items are retrieved. This work considers a new paradigm for the processes used to create and manage controlled vocabularies, since a high degree of integration of these instruments is currently required for the management of data and digital content, which implies the application of controlled vocabularies in logical data analysis processes. For this reason, a more dynamic approach is required, being closer to the design of ontologies than to the creation of traditional documentary languages.

Keywords: *Wikidata*; Ontologies; Controlled vocabularies; Knowledge organization.

1. Introducción

La elaboración de vocabularios controlados supone una de las tareas estratégicas de los profesionales de la información (**Stuart**, 2016, p. 21-23). La digitalización de los procesos de gestión en todos los niveles ha conllevado la presencia de datos estructurados, metadatos descriptivos y todo tipo de contenidos nacidos digitalmente en una gran diversidad de ámbitos y contextos. La experiencia nos muestra que se precisan instrumentos adecuados e interoperables para la correcta descripción y organización de datos y recursos. En consecuencia, la función para la que se idearon tesauros, clasificaciones, encabezamientos de materia y vocabularios controlados en general sigue siendo necesaria. Sin embargo, el escenario de uso ha variado considerablemente desde la concepción de estas herramientas. Ya no es suficiente aplicarlos en los procesos de indización conceptual ya sea manual o asistida. Incluso la actualización de los tesauros a nivel normativo que trajo ISO-25964 (*ISO*, 2011; 2013) se remonta a 10 años. El entorno en el que nos desenvolvemos ya no es el de colecciones con un alto control del flujo de incorporación de nuevos documentos. Poco a poco la rapidez en la publicación y consumo de grandes volúmenes de datos y contenidos ha socavado los fundamentos pragmáticos de los vocabularios controlados. **Dextre-Clarke** (2019) afirma que la vigencia de los tesauros depende en gran medida de la aplicación de tecnologías y su integración en el contexto *linked open data*.

Sin embargo, el despliegue eficiente de estos instrumentos encuentra grandes obstáculos en un mundo en el que los datos, la información y los contenidos se entrelazan en un escenario altamente dinámico. Posiblemente las dos principales limitaciones podrían resumirse en:

- Las limitaciones de la semántica y la formalización que impiden su aplicación en entornos en los que se precise su aplicación en procesos descriptivos avanzados o con un alto nivel de abstracción lógica.
- Su carácter altamente estático que limita su actualización para adaptarse a contextos en los que se precisa la incorporación rápida de nuevos conceptos, etiquetas y relaciones.

Y es que en algunos contextos las aproximaciones tradicionales de vocabularios controlados, organizados mediante estructuras jerárquicas y asociativas puede ser suficiente. Pero en otros, el enfoque clásico comienza a chirriar cuando los engranajes entre objetos de información y vocabularios comienzan a girar en un sistema de información. Tal vez los tesauros, clasificaciones, autoridades, encabezamientos de materia, etc. estén en un punto en el que deban redefinirse aprovechando la oportunidad que supone el desarrollo de soluciones semánticas, tanto conceptuales como tecnológicas.

"El uso efectivo de vocabularios controlados en un escenario digital requiere su adaptación a procesos descriptivos avanzados y a contextos que precisan de una actualización constante"

2. Entendiendo *Wikidata*: ítems, propiedades, declaraciones, afirmaciones, calificadores, referencias, rankings

Wikidata es un grafo de conocimiento basado en un modelo de datos propio que es compatible con RDF. El ítem es el elemento principal de *Wikidata*, posee un identificador único y su descripción se realiza mediante propiedades. En unos casos las propiedades definen relaciones entre ítems, mientras que en otros las descripciones se refieren a valores literales. En ambos tipos de propiedades la semántica está bien definida. Cada ítem tiene su propia página con las declaraciones y correspondientes afirmaciones de datos factuales que los describen. Asimismo, las propiedades también disponen de su propia página donde pueden consultarse las características que las definen, propiedades inversas, ámbito de uso, etc.

Uno de los aspectos más complejos de *Wikidata* es el uso de calificadores, referencias y rankings.

Los calificadores permiten realizar afirmaciones sobre una determinada afirmación. Por ejemplo, en el ítem correspondiente a España (con identificador Q29 y accesible a través de *http://www.wikidata.org*) la declaración que contiene las diferentes afirmaciones del número de habitantes de España (propiedad P1082, población) indica el punto en el tiempo al que se refiere dicho dato.
https://www.wikidata.org/wiki/Property:P1082

Igualmente, el ranking permite especificar cuál de las afirmaciones que hacen uso de una misma propiedad es la preferente. Siguiendo con el ejemplo: de las 58 afirmaciones sobre la población de España, la que está seleccionada como preferente es la que tiene asociado el calificador que indica que el dato se refiere a 2018.

Las afirmaciones también pueden tener asociadas referencias a fuentes de información para verificar la validez de los datos.

Las calificaciones, ranking y referencias se definen utilizando un mecanismo similar a la reificación RDF, lo que refuerza la compatibilidad de *Wikidata* con este modelo de datos y por lo tanto su representación en el contexto *linked data*.

3. Explorando la selva del conocimiento de *Wikidata*

Cuando se trabaja con ontologías u otro tipo de vocabularios controlados es frecuente definir clases para agrupar individuos o elementos. Esto también tiene lugar en *Wikidata*, pero con una peculiaridad: no existen clases definidas de forma explícita y diferenciada, sino que dicho papel lo desempeñan algunos ítems en función de las relaciones que tienen con otros ítems.

Además, es frecuente encontrar estructuras en las que la asociación es temática, funcional, por facetas, etc. Es decir: no existe un catálogo de propiedades que permita identificar los ítems que pertenecen a una u otra clase o temática. Aparentemente, se trata de una organización regida por el caos debido a la diversidad de criterio de los editores. La propiedad P31 (instancia de) es la más apropiada para asociar un ítem con su clase. Sin embargo, también se utilizan otras como:

- P361 (forma parte de),
  https://www.wikidata.org/wiki/Property:P361
- P1269 (faceta de),
  https://www.wikidata.org/wiki/Property:P1269
- P921 (tema principal), etc.
  https://www.wikidata.org/wiki/Property:P921

Entonces ¿por dónde podemos comenzar la exploración de *Wikidata* en relación con un dominio de conocimiento o tema? El primer paso podría consistir en encontrar algún ítem que, en un principio, de forma prospectiva, desempeñe el punto de partida para identificar el dominio que queremos explorar. Podrían usarse los siguientes ítems:

- Q634: Planetas
  https://www.wikidata.org/wiki/Q634
- Q34726: Mitología griega
  https://www.wikidata.org/wiki/Q34726
- Q81738: Legendarium de Tolkien
  https://www.wikidata.org/wiki/Q81738
- Q18043309: Universo narrativo de Star Trek / Q1092: Star Trek
  https://www.wikidata.org/wiki/Q18043309
  https://www.wikidata.org/wiki/Q1092

A continuación, es necesario averiguar el tipo de relaciones que establecen con el ítem que identifica el dominio o tema con el resto de ítems de *Wikidata*. Siguiendo el ejemplo anterior, se podría ejecutar en WDQS (W*ikidata Query S*ervice) la siguiente consulta *Sparql* genérica. Debe sustituirse "ITEM_DOMINIO" por el identificador Q correspondiente al ítem de dominio. La consulta puede combinar varios ítems de dominio separando con espacios en blanco los diferentes identificadores.

```
SELECT DISTINCT ?propiedadLabel ?p ?dominioLabel (COUNT(?sujeto) AS ?num) WHERE
{
  VALUES ?dominio {wd:ITEM_DOMINIO}
  ?sujeto ?propiedad_dominio ?dominio;
    ?p ?objeto.
  ?propiedad wikibase:directClaim ?p.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "es","en".
  }
}
GROUP BY ?propiedadLabel ?p ?dominioLabel
```

La consulta anterior devuelve todas las propiedades utilizadas entre todos los ítems de *Wikidata* que están vinculadas con el ítem (o ítems) seleccionado para referirnos al dominio o tema. Esta exploración inicial es crucial para decidir qué ítem de dominio se utilizará para recuperar todos los ítems que contiene *Wikidata* sobre un determinado tema. En la figura 1 puede verse el mapa de calor correspondiente a las propiedades utilizadas para vincular ítems de *Wikidata* con los ítems de dominio utilizados en el ejemplo.

El gráfico utiliza una escala logarítmica en la intensidad de color debido a la gran diferencia de uso de unas u otras propiedades, pero puede verse la gran heterogeneidad existente en función de la temática sobre la que se realiza la búsqueda.

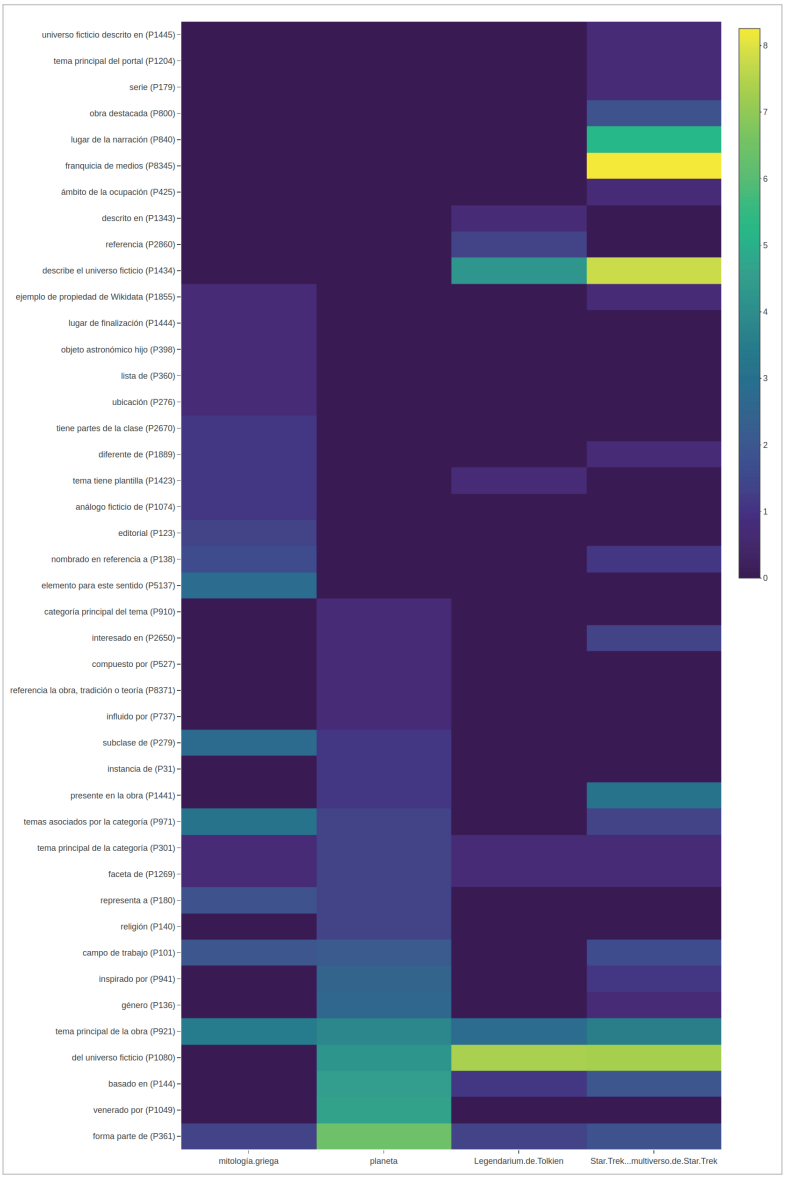

Figura 1: Mapa de calor de las propiedades utilizadas en *Wikidata* para describir algunos dominios de conocimiento.

4. Clases… malditas clases

El paso anterior permite localizar los ítems correspondientes a un tema. Ahora es el momento de recuperar la taxonomía de clases y subclases. Este paso resulta de utilidad para comprender la estructura de organización del conocimiento que la comunidad de *Wikidata* ha definido en un determinado dominio. Al inicio de la sección anterior de esta nota se ha mencionado el modo en el que funciona el concepto de clase en *Wikidata*. También se ha mencionado la heterogeneidad de criterios a la hora de utilizar una u otra propiedad para jerarquizar los ítems que desempeñan el papel de clase.

No hay una receta mágica para recuperar la taxonomía de clases asociadas a un dominio. El principal problema de recuperar esta información es la escalabilidad, ya que este tipo de consultas son complejas y pueden generar tiempos de respuesta altos o directamente un error de "*timeout*". Una consulta que recuperaría los ítems que desempeñan el papel de clase respecto a los ítems de un dominio, los ítems jerárquicamente superior o superclases, las propiedades que los vinculan y los ítems definidos como instancias de las clases mediante la propiedad P31 sería la siguiente:

```
SELECT DISTINCT ?clase ?claseLabel ?propiedad_clase_subclase ?superclase ?superclaseLabel (COUNT(?sujeto) AS ?num) WHERE
  {
    VALUES ?dominio {wd:ITEM_DOMINIO}
    VALUES ?propiedad_clase_subclase {
      wdt:P279 wdt:P31 wdt:P361 wdt:P1269
    }
    ?sujeto ?propiedad_dominio ?dominio;
      wdt:P31 ?clase.
    ?clase ?propiedad_clase_subclase ?superclase.
    SERVICE wikibase:label {
      bd:serviceParam wikibase:language "es","en".
    }
  }
  GROUP BY ?clase ?claseLabel ?propiedad_clase_subclase
        ?superclase ?superclaseLabel
```

La consulta anterior, tal y como muestra la figura 2, permite recuperar una serie de datos para analizar cuantos ítems del dominio son instancia de una clase.

A partir de ese número es posible realizar una selección de clases y superclases para poder definir los criterios que permitan construir una taxonomía para organizar nuestra ontología. La taxonomía clases también puede aplicarse para identificar y recuperar los ítems que son instancias de las mismas para definir los individuos de las ontologías o los elementos de los vocabularios. Evidentemente esta información debe ser revisada y sometida a un proceso de selección y reorganización según los fines del vocabulario u ontología que se desee crear.

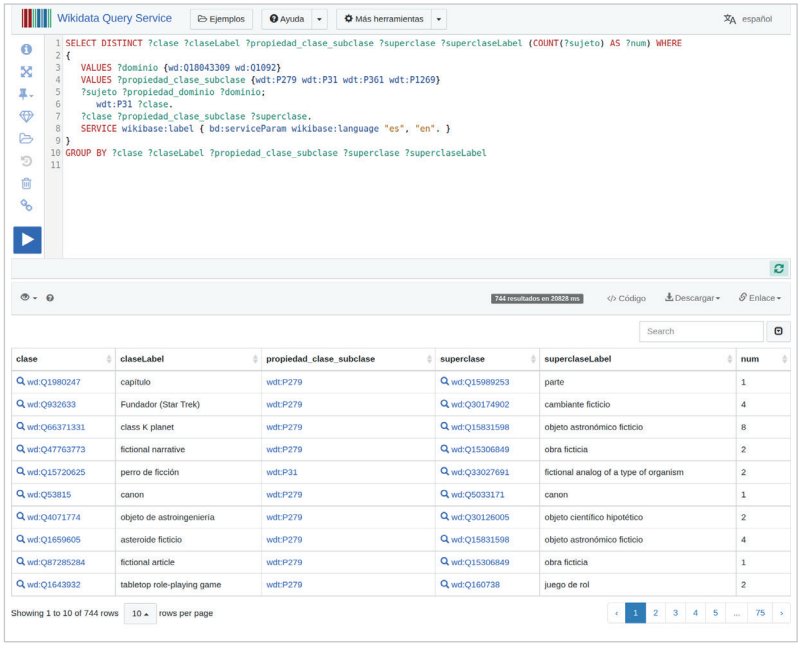

Figura 2: Clases de las que son instancias los ítems del dominio sobre *Star Trek*. La última columna muestra el número de ítems asociados a dichas clases.

4. Esta propiedad sí… esta propiedad no

En *Wikidata* también puede realizarse la búsqueda de las propiedades utilizadas para describir los ítems de un dominio de conocimiento. A partir del análisis de los resultados se pueden identificar aquellas más adecuadas para nuestro vocabulario u ontología.

En el momento de difundir esta nota *Wikidata* tiene 9001 propiedades disponibles. Sin embargo, no todas se aplican en todos los ítems. La propiedad P1604 (nivel de bioseguridad: https://www.wikidata.org/wiki/Property:P1604) puede ser útil para describir el ítem sobre el SARS-CoV-2 (Q82069695: https://www.wikidata.org/wiki/Q82069695), pero es de difícil aplicación para describir el ítem Q177329 correspondiente a Frodo Bolsón (https://www.wikidata.org/wiki/Q177329).

Nuevamente esto se puede hacer de un modo relativamente sencillo con *Sparql*:

```
SELECT DISTINCT ?propiedadLabel ?p ?dominioLabel (COUNT(?sujeto) AS ?num) WHERE
{
  VALUES ?dominio {wd:ITEM_DOMINIO}
  ?sujeto ?propiedad_dominio ?dominio;
    ?p ?objeto.
  ?propiedad wikibase:directClaim ?p.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language “es”,”en”.
  }
}
GROUP BY ?propiedadLabel ?p ?dominioLabel
```

Para algunas propiedades de *Wikidata* se han definido las correspondientes equivalencias con otras ontologías y vocabularios de metadatos. Esto resulta de gran utilidad al pensar en las posibilidades de interoperabilidad e integración en el ecosistema *linked data* de la ontología o vocabulario controlado que estemos diseñando. Sin embargo, como puede verse en la figura 3, todavía queda mucho camino por recorrer por parte de la comunidad de *Wikidata*.

“No existe una única fórmula o camino para la obtención de los datos en *Wikidata*. Las estrategias varían según el dominio de conocimiento o el volumen de datos existente sobre el mismo”

Puede observarse en ambos gráficos que tanto para las propiedades que describen los dominios de conocimiento sobre Mitología griega (arriba) como el de *Star Trek* (abajo) todavía no se han definido propiedades equivalentes. Cabe destacar el mapeado existente entre *Wikidata* y otras iniciativas externas como *Schema.org*, *Dbpedia* y *GND* y los vocabularios de *Getty*.

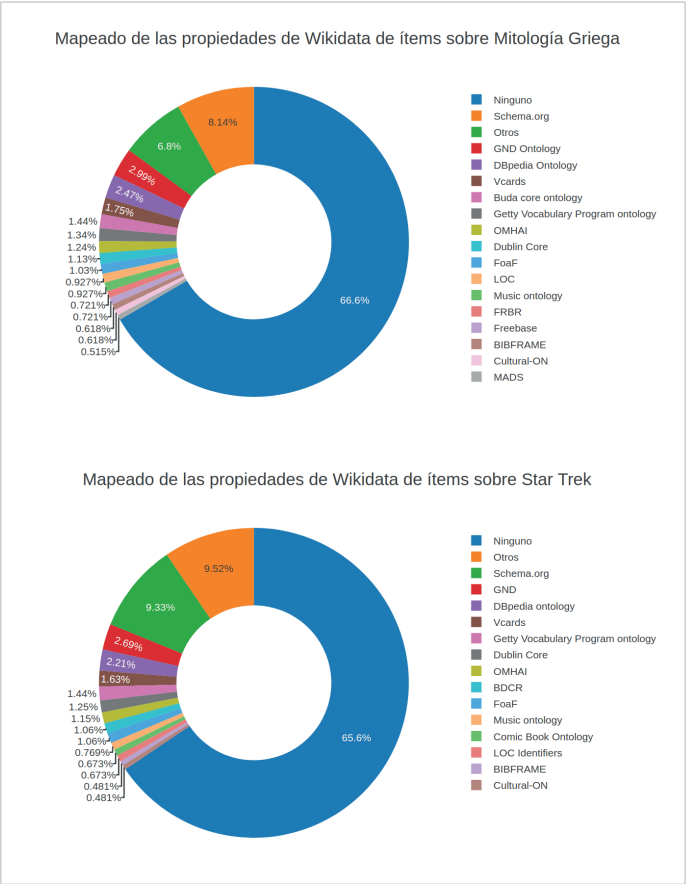

Figura 3: Porcentaje de propiedades de *Wikidata*, utilizadas para describir los dominios sobre Mitología griega y *Star Trek*, mapeadas con propiedades de vocabularios y ontologías externas.

5. Conclusiones: una reflexión del pasado al futuro

Wikidata es una valiosa fuente de información en el ámbito de las ontologías y vocabularios controlados. Evidentemente los principales problemas para su aplicación en este campo están relacionados con la calidad de los datos. Con el paso del tiempo la comunidad se está organizando para asegurar la veracidad y coherencia de los datos almacenados en *Wikidata*. Pese a todo, es una herramienta de gran utilidad para la obtención de terminología multilingüe para el etiquetado de individuos/conceptos, definición de taxonomías de clases y la definición y establecimiento de relaciones semánticas.

No existe una única fórmula o camino para la obtención de los datos. Las estrategias pueden variar en función del dominio de conocimiento en el que se trabaje o el volumen de datos existente sobre el mismo.

En todo caso *Wikidata* nos acerca a una reflexión necesaria sobre el papel, naturaleza y estructura de los vocabularios controlados que tradicionalmente se denominaban lenguajes documentales. En este punto merece la pena recordar y releer los trabajos de **García-Jiménez** (2004) y **García-Marco** (2007). Pasar a un paradigma dominado por las ontologías supone algo más que la mera aplicación del modelo de datos de *SKOS*. Tal vez nos encontremos en un momento en el que se precise la reinvención en donde las relaciones jerárquicas poco definidas o el cajón de sastre de las relaciones asociativas deba enriquecerse con propiedades con una semántica más definida. En ese momento los vocabularios controlados darán el gran paso, aparentemente inevitable, de unificación con las ontologías y que permitirá su aplicación en procesos que requieran una mayor formalización lógica del proceso de organización del conocimiento.

6. Referencias

Dextre-Clarke, Stella G. (2019). "The information retrieval thesaurus". *Knowledge organization*, p. 46, n. 6, pp. 439-459. *https://www.nomos-elibrary.de/10.5771/0943-7444-2019-6-438/*

García-Jiménez, Antonio (2004). "Instrumentos de representación del conocimiento: tesauros versus ontologías". *Anales de documentación*, v. 7, pp. 79–95. *https://revistas.um.es/analesdoc/article/view/1691*

García-Marco, Francisco-Javier (2007). "Ontologías y organización del conocimiento retos y oportunidades para el profesional de la información". *El profesional de la información*, v. 16, n. 6, pp. 541-550. *https://doi.org/10.3145/epi.2007.nov.01*

ISO (2011). *ISO 25964-1:2011. Information and documentation. Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval*. Ginebra: International Society for Standardization.

ISO (2013). *ISO 25964-2:2013. Information and documentation. Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies*. Ginebra: International Society for Standardization.

Stuart, David (2016). *Practical ontologies for information professionals*. Londres, Facet publishing. ISBN: 978 1 78330 152 2