CONSULTORA DE CIENCIAS DE LA INFORMACIÓN
BUENOS AIRES
ARGENTINA

Serie

DOCUMENTOS DE TRABAJO

Área: Procesos Técnicos

Proyecto para el desarrollo de un vocabulario controlado de datos abiertos enlazados en el área de Ciencias de los Polímeros

Lic. Marcelo de la Puente

julio 2020

N° 076

ISSN 1852 - 6411

Editor: Patricia Allendez Sullivan. Asistente Editorial: Analía Bedrosian

Puente, Marcelo de la

Proyecto para el desarrollo de un vocabulario controlado de datos abiertos enlazados en el área de ciencias de los polímeros. Buenos Aires: Consultora de Ciencias de la Información, 2020.

ISSN 1852 – 6411

1. Vocabulario Controlado 2. Datos Enlazados 3. Ontología 4. Ciencia de los Polímeros
I. Título

Resumen

El presente trabajo consiste en el desarrollo de un vocabulario controlado para aplicar en el área de las Ciencias de los Polímeros. Para ello se analizarán clasificaciones y vocabularios existentes. En todo momento se tendrá en cuenta las características semánticas de los vocabularios existentes y la manera de adaptarlos para recuperar información de recursos web en un entorno abierto y compartido, vinculándolos a otros lenguajes documentales en distintos idiomas, lo que implica la reconversión de esos vocabularios a formatos compatibles.

Introducción

La gestión eficiente de la información contenida en recursos ha sido siempre una labor fundamental de la actividad de los profesionales de la información, pero con el fenómeno de la así llamada explosión de la información, que se ha dado a partir de la automatización de los procesos de gestión de la misma en unidades de información, con la proliferación de múltiples formatos electrónicos, el crecimiento de las publicaciones científicas y en particular, con la aparición de Internet y el crecimiento exponencial de la información en la misma, lleva a nuevos retos y desafíos profesionales para lograr esta tarea profesional básica.

Tradicionalmente en bibliotecas, archivos y centros de documentación los vocabularios controlados como las clasificaciones, las listas de encabezamiento de materia, los tesauros han permitido organizar el conocimiento en categorías semánticas definidas, tanto en el almacenamiento como en la recuperación de información, sean vocabularios precoordinados, en el caso de encabezamientos de materia, o postcoordinados en el caso de los tesauros, pero dada la complejidad y la cantidad monumental de información de los recursos digitales, así como el predominio en la web de texto libre no estructurado hacen menos eficientes estos formatos tradicionales bibliotecario en la categorización semántica y plantea nuevos desafíos.

El objetivo de este trabajo es analizar las clasificaciones y vocabularios existentes en el campo de la ciencia de los polímeros y temas relacionados con el propósito de proponer elementos o principios generales para la realización o la conversión de vocabularios preexistentes a un vocabulario de datos abiertos enlazados o de datos enlazados, que permita la recuperación semántica de información a partir de un gran número de fuentes, recursos web, bases de datos, etc.

Se deben tener en cuenta, en primer lugar, las características semánticas de los vocabularios existentes y como adaptarlos a la recuperación de información de recursos web en un entorno abierto y compartido, permitiendo la vinculación con otros lenguajes documentales en diferentes idiomas, lo cual implicaría la reconversión de estos vocabularios a formatos compatibles y que puedan ser extendidos para su reutilización y el uso de aplicaciones autónomas.

Se analizarán a continuación las características de tres vocabularios: la clasificación facetada para la Ciencia de los Polímeros de RAPRA, el tesauro de la misma institución y la ontología de la base PolyInfo, para a continuación enumerar las características que deben cumplir los vocabularios de datos abiertos enlazados según el W3C y que herramientas pueden utilizarse para lograr esa conversión.

Vamos a comenzar

La idea general, similar a otros proyectos realizados sobre el tema, seria realizar un análisis preliminar sobre los vocabularios controlados del sector, el tipo de relaciones semánticas que presentan y la posibilidad de lograr una mayor especificación y profundidad semántica de las mismas, con miras a su reutilización por parte de aplicaciones automatizadas y agentes de software en la web, tal como ocurre en las ontologías y taxonomías de recursos digitales, ya que a diferencia de lenguajes documentales como listas de encabezamientos de

materia y tesauros, concebidos para ser utilizados por seres humanos, solo se especifican una relativamente pequeña cantidad de relaciones semánticas, tales como las relaciones de equivalencia (sinonimia) , las relaciones jerárquicas (género-especie y todo-parte) y las relaciones asociativas (términos relacionados).

En cambio, en estructuras como las ontologías, por ejemplo, las relaciones semánticas alcanzan un nivel mucho mayor de profundidad semántica, ya que deben representarse en lenguajes formales para que sean pasibles de utilizadas por agentes de software y programas de múltiples aplicaciones web. Se deben poder realizar las descripciones adecuadas, que permitan incorporar las descripciones multilingües para poder aumentar la capacidad de recuperación de información en múltiples plataformas, adoptar formatos extendidos para su libre reutilización y gestión por aplicaciones autónomas.

En el caso de los tesauros para que sea posible su utilización en un entorno de datos enlazados entre recursos en la web se requiere que sean diseñados conforme a normas como ISO 25964, que permiten actualizar la gestión de datos en un contexto distribuido, en cuanto a modelo de datos, su estructura conceptual en la representación léxica, definición de varios niveles de grupos semánticos de conceptos y de relaciones de equivalencia, entre otros aspectos.

Los datos en linked data o datos enlazados permiten la interconexión de datos, lo que implica previamente, su descripción adecuada por medio de metadatos y la normalización de identificadores de estos a través de lenguajes documentales y ontologías, esta normalización es un condición previa y necesaria para lograr la interoperabilidad consistente de los datos.

Los principios básicos de los datos abiertos enlazados según Tim Bernes Lee, son:

1- Usar URI o Identificadores únicos de Recursos para nombrar cosas
2- Usar URI HTTP que sean interpretados por humanos y máquinas
3- Proveer información útil a través de un estándar o norma para representar los datos en la web, como RDF, OWL, SPARQL (lenguaje de consulta para la obtención de información a partir de conjuntos de datos en RDF)
4- Crear Links entre URI para favorecer el proceso de descubrimiento de objetos

El patrón de uso recomendado es:
http://(dominio)/(tipo)/(concepto)/(referencia)

Siendo el dominio el sector desde el que se elabora el URI y el host si es relevante. El tipo de recurso se idéntica con un código semántico, por ejemplo, documento (doc), concepto es el conjunto referido por la URI que puede ser una colección de objetos, recursos, esquema, etc. La referencia específica los conceptos

Un protocolo de Linked data implica el facilitar el acceso a los datos y su descubrimiento mediante el uso de tecnologías semánticas de búsqueda, lo que implica metadatos con anotaciones semánticas de los recursos web permitiendo la utilización de grandes conjuntos de datos, lo que permite pensar a la web como una gran base de datos, una red interconectada mediante enlaces semánticos y orientada al procesamiento automatizado de los mismos. Se puede considerar que existen URIs que identifican conceptos y por otro lado los que identifican recursos web, por lo que cada recurso puede ser identificado por tres URIs diferentes: el enlace al concepto en abstracto, hacia el objeto o documento legible por humanos (en HTML) y a la propiedad o descripción semántica (en RDF) legible por programas.

Puede decirse que en el nivel semántico moverse desde los términos hacia los conceptos con URIs hace la actualización de la información más fácil de

implementar. Deben utilizarse herramientas y estándares que permitan una mayor especificidad semántica de los vocabularios y definirlas en estándares apropiados para el entorno web de aplicaciones abiertas. Ya existen en la web vocabularios como los tesauros digitales, por ejemplo que permiten una descripción más exhaustiva del contenido temático permitiendo de esta forma, la recuperación por diferentes aspectos, puntos de vista o facetas, permite controlar sinónimos, homónimos y cuasinónimos, es decir términos que son conceptos afines que pueden remitir a los sinónimos, se pueden controlar las relaciones de sinonimia, jerárquicas (de género-especie y partitivas) y asociativas, permitiendo añadir funciones hipertextuales, hipermediales y objetos mulitimedia. Pero para un entorno web se requiere una mayor especificación semántica, como las que pueden brindar las *Ontologías.*

Las ontologías son especificaciones formales, explicitas de una conceptualización de un dominio temático determinado, en ellas se definen los conceptos o clases principales de ese dominio, algunas de sus propiedades y la relación entre los conceptos, se componen en esencia, *clases* o subclases, *slots* que definen los roles, propiedades o características de cada concepto ya sean propiedades extrínsecas o intrínsecas, las *facetas*, que definen los tipos de valores o valores permitidos para las propiedades y finalmente las *instancias* u ocurrencias de una clase dada.

De esta manera, las ontologías es decir permiten utilizar varias categorías gramaticales, como adjetivos, verbos y adverbios, no solo sustantivos como es el caso de los tesauros, son más flexibles para el empleo de conceptos complejos como los utilizados por los usuarios en búsqueda, ya que los términos que utilizan están más cercanos al lenguaje natural, una condición necesaria para las búsquedas en la web, permiten un nivel más profundo de descripción del vocabulario, una especificación semántica de las relaciones entre conceptos, en particular para las relaciones jerárquicas, clase subclase y para las relaciones cruzadas o asociativas muy superior y más específicamente que un tesauro y se

construyen en lenguaje formal legible por aplicaciones web, lo cual permite la interoperabilidad entre diferentes, por lo que permiten una mayor reusabilidad en sistemas y aplicaciones, ya que describen formalmente objetos, propiedades y relaciones. Las ontologías se basan en RDF, el cual se base en el modelo del triplete de datos como el sujeto o recurso, el predicado u etiqueta preferida y el objeto, elementos que permiten definir con precisión las clases y subclases, las propiedades y las limitaciones a los valores permitidos de las mismas en una ontología.

Los lenguajes de etiquetado o de codificación que permiten construirlas son varios, RDF, OWL, DAML + OIL, etc., RDF (Resource Description Framework), es un lenguaje de etiquetado de recursos web que genera objetos de información compuestos por elementos codificados con el uso de etiquetas de a pares con atributos que funcionan como contenedores de datos, independientemente de su formato. RDF describe los recursos documentales mediante un conjunto de propiedades. Define un *recurso de información* como un objeto que puede ser identificado por un URI (Identificador de Recurso Uniforme). Las *propiedades*, tienen un tipo o valor que puede ser elemental u otro recurso. La colección de esas propiedades referidas a un recurso se denomina *descripción*, que puede ser también elemental o apuntar a otro recurso.

En síntesis, un recurso puede tener como descripción elemental parte de su texto y como descripción a otro recurso, identificadores como el nombre del autor, que a su vez tiene otros identificadores, es decir, los recursos pueden ser páginas web o cosas como personas y objetos del mundo real o conceptual. Las propiedades son las características relevantes de los recursos, como por ej., el autor o el idioma.

RDF permite el uso de dos lenguajes de ontologías que permiten un modelado adecuado: RDFs y OWL.

RDFs permite un modelado de datos apropiado para una representación de datos en una ontología, establecer relaciones semánticas entre diferentes elementos para lograr definir las características de los dominios temáticos y rangos de las diferentes propiedades. Describe los recursos clasificándolos en clases y sus elementos, las instancias, permitiendo la organización de los diferentes elementos con etiquetas como: rdf resource, rdf Class,rdf Type of Date para el tipo de datos permitidos, por ejemplo., :rdf Property para las propiedades, entre otra que establece las relaciones entre el sujeto y el objeto y permite organizar los datos en un triplete o conjunto de tres datos (sujeto, predicado y objeto).

OWL(Ontology Web Language) o lenguaje de ontologías permite un procesamiento más específico de los datos, adecuado para la formación de una ontología. Se estructura en tres ejes fundamentales: Los *axiomas* o enunciados básicos, las *entidades* o elementos que hacen referencia a objetos o entidades reales y las *expresiones*, que son combinaciones de entidades formando complejas descripciones de axiomas. Propiedades de objetos, de tipos de datos y propiedades de anotaciones semánticas, lo cual es un principio fundamental para que los sistemas automatizados que operan con ellas puedan efectuar inferencias sobre las relaciones entre los diferentes conceptos de estas y llegar de esta forma, al descubrimiento de nuevos recursos.

Los principales elementos de OWL son: *clases e instancias* que representan a grupos de individuos cuya unión se establece por formar parte de un concepto determinado, unas clases pueden abarcar jerárquicamente unas a otras, es decir, serían clases y subclases implicadas con relaciones jerárquicas, de pertenencia por medio de la propiedad transitiva (si B es subclase de A y C es subclase de B, entonces C es subclase de A), relación de equivalencia entre clases que se refieren al mismo conjunto de individuos, si dos individuos pertenecen a clases excluyentes, se expresa mediante la clase de disyunción, la relaciones entre individuos se expresan por propiedades, en la jerarquía de

propiedades se verifica el mismo proceso de las equivalencias, la inferencia de conocimiento se realiza en individuos conectando por las propiedades, ello permite extraer el dominio y el rango de valores permitidos, etc. Se analizará más en profundidad el lenguaje OWL en el siguiente apartado.

Todas estas estructuras son necesarias para el desarrollo de la Web Semántica, denominada así porque permitiría efectuar un procesamiento de la información más profundo, con la capacidad de procesar la información procedentes de diferentes recursos y sitios web, discriminar aquellos que puedan ser más adecuados en relación a un requerimiento, deducir o inferir información no registrada y tomar decisiones con cierto grado de autonomía, lo que implica el establecimiento de estándares comunes de comunicación y el empleo de enlaces conceptuales entre las palabras y la posibilidad de saltar a través de conceptos, documentos, sistemas y autores, lo cual le facilita a los programas el procesamiento semántico de la información.

Este procesamiento está más basado en un análisis semántico más que en la coocurrencia estadística en la frecuencia de aparición de términos de búsqueda. Un entorno de linked data o datos enlazados en este contexto permitiría el establecimiento de relaciones entre conceptos y recursos, generando de esta forma, una red semántica formada por documentos interconectados, en la que los documentos pueden considerarse como parte o extensión de los conceptos, lo que permitiría visualizar campos semánticos integrados por recursos.

SKOS (Simple Knowledge Organization Systems)

Otra estructura utilizada para la representación de vocabularios para el uso en un entorno de datos enlazados es *SKOS* o *Simple Knowledge Organization Systems*, que es considerado un lenguaje ontológico más general de amplio espectro en comparación con OWL, lo que permite un modelo de orden mas general de los conceptos de un dominio temático y sus relaciones. Establece la

idea de un concepto (*skos:Concept*), según la norma ISO 25964, separada de otros conjuntos de agrupación como la colección (*skos:Collection)*, etiqueta (*skosxl:Label*.), categoría semántica (*skos:ConceptScheme)*, entre otros. Colección no se encuentra organizada en el mismo nivel sintáctico que las categorías semánticas y no puede establecer relaciones semánticas con otras estructuras del modelo, pero si puede contener conceptos agrupados por campos semánticos sin una unidad estructural, lo que lo hace adecuado para representar dominios específicos de una materia más general en la colección. SKOS es un lenguaje basado en RDF para la representación de esquemas de conceptos como las clasificaciones, las taxonomías, los tesauros, etc.

En el contexto de linked data o de datos enlazados, es su representación en la web de forma de permitir su interoperabilidad y su reutilización. Se compone, entonces de Conceptos (*skos:Concept*),, etiquetas léxicas, siguiendo los elementos prefigurados para la norma que prefigura la estructura de los tesauros: la etiqueta preferente (*skos:prefLabel)*, siendo únicas para cada lengua, lo que permite la equivalencia multilingüe, las etiquetas alternativas o no preferidas (*skos:prefLabel)*, las relaciones jerárquicas, los términos generales (*skos:broader),* términos específicos (*skos:narrower*) y las relaciones asociativas, los términos relacionados *(skos:related*), permite expresar la transitividad de propiedades en la jerarquía para mejorar su aplicación en las relaciones automáticas (*skos:broaderTransitive* y *skos:narrowerTransitive)*.

En lo concerniente a las relaciones semánticas dentro de un dominio temático dado, se puede utilizar la etiqueta skos*:ConceptScheme* que permite la estructuración de campos semánticos más granulares. Los conceptos declaran su pertenencia a un esquema determinado mediante la propiedad *skos:inScheme*, pudiendo además cada concepto pertenecer a más de un esquema, lo que permite la representación de relaciones multijerárquicas. El concepto principal o raíz de cada jerarquía es el *skos:hasTopConcept*, que pueden declarar su pertenencia a un esquema determinado mediante la

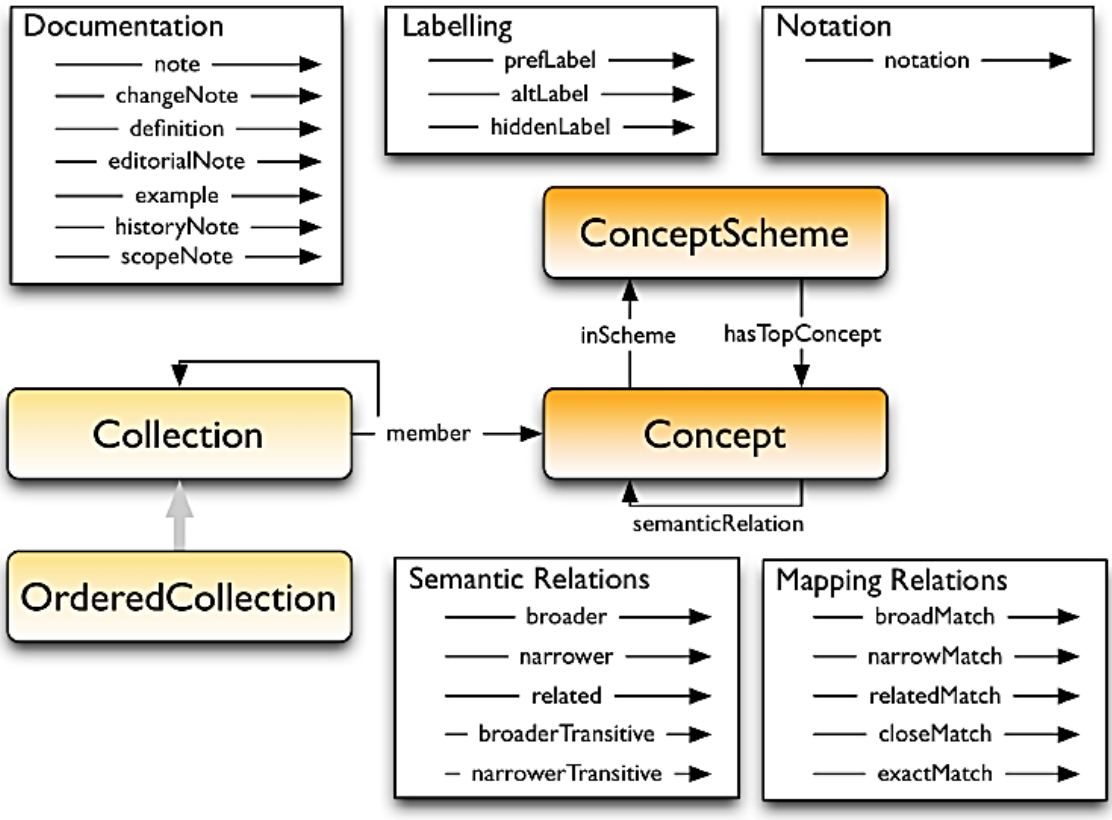
propiedad *skos:topConceptOf*. La etiqueta colección *skos:collection* permite la agrupación de grupos de conceptos identificados por una etiqueta. La pertenencia de un concepto a una colección se representa por la etiqueta *skos:member*.

Las listas ordenadas de conceptos se representan mediante las etiquetas *skos:OrderedCollection y skos:memberlist.*

SKOS derivado en primera instancia de RDF, permite la representación de diferentes lenguajes documentales como taxonomías, esquemas de clasificación, listas de encabezamientos de materias, tesauros, etc. , en el contexto de los datos enlazados, permite representarlos en un entorno web permitiendo la interoperabilidad y la reutilización de la mismos, mediante la etiqueta *skosxl:Label:*, que permite la identificación de relaciones entre entidades léxicas cuya instancia puede ser un recurso web identificado por una URI determinada. Además, el uso de las etiquetas preferentes, no preferentes y ocultas, *skosxl:prefLabel skosxl:altLabe skosxl:hiddenLabel* permite ocultar esas etiquetas para la parte del interfaz humano, manteniéndolas accesibles para el procesamiento automático de las mismas, lo que proporciona una base para poder enlazar conceptos vinculados.

Por otro lado, una de las limitaciones del modelo SKOS es que no definen las relaciones inversas al aplicar las etiquetas de esquema y miembros *skos:inScheme skos:menber*, por lo que se necesita marcado semántico adicional con RDF para poder superar este problema y tampoco permiten definir los puntos de acceso a la consulta de conjuntos de conceptos, lo que es relevante para la construcción de microtesauros ,por ejemplo, por lo que se requiere agregar conjuntos de metadatos adicionales para permitir la identificación. SKOS permite además el mapeo entre distintos vocabularios mediante la equivalencia aproximada de distintas propiedades como *skos:exactMatch skos:closeMatch* o para las relaciones jerárquicas y

asociativas skos:broadMatch, *skos:narrowMatch skos:relatedMatch*. Pero no contempla en su modelo la representación explicita de la relación concepto recurso, por lo que se requiere una actualización posterior.



SKOS permite representar las relaciones semánticas de la mayoría de los lenguajes documentales, pero se adapta principalmente a los tesauros, sobre todo a los especializados en un dominio temático particular, ya que permite la representación de microtesauros.

Tomando por ejemplo la entrada de un tesauro del ámbito de ciencia y técnica, correspondiente a polímeros, puede observarse que el formato SKOS permite representar todas sus relaciones semánticas

PLASTICOS

- *UP Materiales plásticos*
- *UP Productos plásticos*
- TG POLIMEROS
- TG PRODUCTOS SISTETICOS
- TG SINTESIS DE COMPUESTOS ORGANICOS
- TE7ESPUMAS PLASTICAS
- TE7GOMAS Y RESINAS SISTETICAS ►
- TE7PAPEL PLASTICO
- TE7PELICULAS DELGADAS ►
- TE7Películas plásticas
- TE7PLASTICOS A BAJAS TEMPERATURAS
- TE7PLASTICOS EN LA CONSTRUCCION
- TE7PLASTICOS REFORZADOS
- TE7POLIESTERES ►
- TE7POLIMERIZACION ►
- TE7POLIMEROS VINILICOS
- TE7RESINAS EPOXI ►
- TE7RESINAS FENOLICAS
- TE7TERMOPLASTICOS
- TE7TUBERIAS DE PLASTICO
- TR ELASTOMEROS
- TR PLASTICIDAD
- EQ Plastics (Término en inglés)

SKOS permite representar tanto las relaciones de equivalencia o de sinonimia, con términos preferidos y no preferidos:

Skos; prefLabel "Plásticos"
Skos: AltLabel "Materiales Plásticos"

Skos: HiddenLabel "Productos Plásticos" (si se quiere que permanezca oculta pero potencialmente utilizable para la recuperación de información)

Relaciones jerárquicas:

Skos: broader "Polímeros" (Término general)

Skos: narrower "Termoplásticos" (Término específico)

Relaciones asociativas:

Skos: related "Elastómeros" (Términos relacionados)

El mismo concepto en otra lengua:

Skos: alt Label "Plastics" @en

OWL (Ontology Web Language)

El lenguaje de ontologías web es un lenguaje diseñado para la construcción y el diseño de ontologías en la web. Permite extraer razonamientos y conclusiones de las declaraciones basándose en la lógica de la validez implícita de las premisas que soportan una declaración o una afirmación. Eso le permite a diferentes aplicaciones realizar inferencias lógicas a partir de las mismas y es uno de los requisitos básicos para la captura automática de conocimiento. Se estructura en tres pilares fundamentales: 1- Los *axiomas* o enunciados básicos, 2- Las *entidades* o elementos que hacen referencia a objetos reales y 3- Las *expresiones* son combinaciones de entidades que forman complejas descripciones de axiomas. Se identifican a los objetos como individuos, las categorías temáticas como clases y las relaciones como propiedades. A su vez, las propiedades se subdividen en: propiedades de objetos que relacionan objetos con otros objetos, propiedades de tipos de datos a los que le asignan determinados valores o rango y las propiedades de anotaciones que codifican la información de la ontología.

Los principales elementos son:

1. Clases e instancias: representación de grupos de individuos que configuran un campo semántico determinado. Se representa por la etiqueta *ClassAssertion*
2. Si una clase abarca a otra más específica, esto se representa por la relación jerárquica Clase-Subclase, lo cual permite la transitividad de las propiedades a lo largo de varios niveles jerárquicos *SubClassOf*
3. Relación de equivalencia entre clases si se refieren al mismo grupo de individuos EquivalentClasses
4. La clase de disyunción representa a individuos que pertenecen a clases mutuamente excluyentes *DisjointClasses*
5. Permite describir las relaciones entre individuos a través de propiedades
6. Permite expresar jerarquías de propiedades de igual forma que en la jerarquía de clases
7. Permite inferir conocimiento a través de las propiedades y el rango y dominio de individuos conectados por las mismas

En conclusión, puede afirmarse que, para construir un vocabulario en un formato compatible con los datos abiertos enlazados en la web, pueden utilizarse tanto el formato SKOS como OWL para la definición de ontologías. SKOS permite el mapping o la concordancia con otros vocabularios y definir colecciones ordenadas y agrupaciones de conceptos, mediante las etiquetas *skos:exactMatch, skos:closeMatch o skos:relatedMatch*, equivalencia exacta, cercana o relacionada, respectivamente, lo que supone la vinculación de conceptos. SKOS permite la inclusión de un concepto en un vocabulario en otro para completarlo. El uso de SKOS en RDF permite obtener documentos en un formato que facilita su lectura por parte de distintas aplicaciones informáticas, así como su intercambio y su publicación en la web, permite crear nuevas estructuras de organización de conocimiento o adaptar las ya existentes en formatos compatibles con el funcionamiento de la web semántica y puede ser utilizado con OWL conjuntamente o en forma independiente, pero se considera que es un paso intermedio entre los formatos no estructurados del texto libre y mulitimedia que se encuentran normalmente en documentos web y un nivel de

formalización y rigor mucho más elevados utilizados en los lenguajes de descripción de ontologías como OWL.

Existen equivalencias entre las clases generales de SKOS y de OWL, para las clases generales, las agrupaciones de conceptos (lista de miembros, etc.), la documentación (notas de alcance, notas históricas, etc.) y las relaciones semánticas entre conceptos (jerárquicos y asociativos). Pero transformar un vocabulario, un sistema de clasificación, un tesauro en una ontología OWL conlleva un gran esfuerzo debido a que la ontología no proporciona un modelo de datos fácilmente aplicable, ya que los vocabularios controlados se han desarrollado sin una lógica formal, pero se pueden llegar a utilizar en un nivel de formalización requerido para un vocabulario controlado particular, preferentemente en la forma de SKOS.

Los elementos del modelo SKOS y también OWL utilizan clases y propiedades. La estructura y la integridad de los datos están definidas por las características lógicas y las relaciones entre las clases y las propiedades, lo que permite organizar esquemas de conceptos.

Estos conceptos se identifican mediante URIs y pueden ser etiquetados en cualquier idioma. Cada concepto puede tener asociadas múltiples etiquetas para un idioma, pero solo una se considera como la etiqueta preferente, el resto son etiquetas alternativas y pueden permanecer ocultas y ser utilizadas en los procesos de indización y de búsqueda automáticos por distintas aplicaciones. Pueden también asignarse a los conceptos códigos de clasificación o de identificación para un esquema conceptual determinado, por lo que también es un formato valido para ser utilizados para un sistema de clasificación especializado, no solo para los tesauros. Finalmente, los conceptos pueden agruparse en colecciones y a la vez pueden etiquetarse y ordenarse en diferentes colecciones.

Lenguajes documentales en el campo de la ciencia de los polímeros

A continuación, se evaluarán algunos de los vocabularios controlados más relevantes en el campo de la Ciencia de los Polímeros y algunos intentos que se han hecho para convertirlos en un formato compatible con su representación para la recuperación semántica de la información.

Los polímeros son grandes moléculas o macromoléculas creadas como resultado de pequeñas unidades constitutivas denominadas monómeros. Según su composición química los polímeros pueden componerse de monómeros idénticos y en este caso se denominan *homopolimeros*, como las poliolefinas, poliestirénicos, polienos, polivinilos y poliacrílicos. Pero si se componen de dos o más unidades químicas diferentes, se denominan *copolimeros*, que pueden ser de distintas características aleatorios, alternados, en bloque, etc. Pueden también ser naturales, como es el caso de la celulosa y las proteínas, por ejemplo, semisintéticos, es decir, naturales con alguna modificación, como el caucho vulcanizado, por ejemplo, o sintéticos, creados en base a polímeros naturales en laboratorios.

En cuanto a la tecnología de procesamiento , y propiedades de los polímeros, pueden ser *termoplásticos*, polímeros con una conexión laxa entre las cadenas atómicas que se ablandan con el calor, luego al enfriarse pueden recalentarse y pueden utilizarse para formar otros objetos *termoestables*, con una estructura cruzada de cadenas que no pueden separarse nuevamente luego de su calentamiento, haciendo al polímero más resistente hasta el punto en que se degrada y *elastómeros*, con una conformación más irregular de las cadenas atómicas, que al sufrir una deformación vuelven a su forma original.
Según su forma, pueden ser lineales o con ramificación en las cadenas de átomos y según su mecanismo de polimerización, pueden ser de adición, en el caso de monómeros que cuentan con varias conexiones por condensación, donde se desarrollan paso a paso a partir de moléculas con bajo peso

molecular. La composición de las macromoléculas puede ser de estructura amorfa, semicristalina o una estructura cristalina liquida.

Todo lenguaje documental especializado en el campo sea clasificación, tesauro u otro tipo de vocabulario, debe dar cuenta de estas posibilidades de clasificación según todos estos puntos de vista.

El esquema de clasificación más importante en el campo de la ciencia de los polímeros es la clasificación RAPRA, creada por la asociación RAPRA (Rubber and Plastic Research Association), una asociación sin fines de lucro de alcance mundial para empresas que producen o comercializan materiales poliméricos que ha elaborado estándares, proyectos de investigación en el sector entre otras actividades por más de 90 años.

La versión original del esquema fue desarrollada por late T.R. Dawson, para el uso del centro de información de la institución cuando tenía su nombre original Research Association of British Rubber Manufacturers (RABRM) en 1937. Se publicó una revisión del código en 1942 y en 1946. A comienzos de la década de los años 60 RAPRA tomó la decisión de hacer una revisión general del esquema de clasificación, lo que llevo a la publicación de un nuevo esquema en 1964-. Posteriormente se fueron publicando adiciones y revisiones que se incorporaron finalmente a la edición publicada en 1994.

La clasificación emplea una notación alfanumérica para la identificación de clases y su ordenamiento por un sistema decimal, con números en orden preferente antes que las letras. Las clases principales del código son:

0 GENERAL
1 ORGANIZACION INDUSTRIAL, ADMINISTRACION Y ECONOMIA
2 MAQUINAS, PLANTAS, ENSAYOS, EQUIPOS E INSTRUMENTOS
3 MATERIAS PRIMAS INCLUYENDO MONÓMEROS

4 POLIMEROS Y RESINAS
5 COMPOSICION DE INGREDIENTES Y SOLVENTES
6 APLICACIONES DE POLIMEROS
7 DISEÑO DE NORMAS Y ESPECIFICACIONES, FUENTES Y PRODUCCION
8 PROCESAMIENTO Y TRATAMIENTO
9 PROPIEDADES Y ENSAYOS

En el caso de la clase de los polímeros de adición 42, se brinda una lista de comonómeros y se construye el número por el monómero que aparece en primer lugar seguido por el número del que aparece después en el orden, por ejemplo,

42C131D12 Butyl rubber
42C21C391D11 Styrene-acrylonitrile-butadiene terpolymer

Un comonómero principal puede denominarse con una letra mayúscula y es seguido por el número de clase correspondiente a ese comonómero, por ejemplo:

42C382A Vinyl chloride copolymers

Se utiliza el colon para unir diferentes aspectos o facetas de clases principales por ejemplo,

Properties of butadiene-styrene copolymers
42D11C21:9
Properties of butadiene-styrene passenger car tyres reinforced with carbon black
42D11C21:51B:6T11:9

Por lo que puede describirse como una clasificación facetada en un campo especializado, sobre la cual posteriormente se desarrolló un tesauro, el *RAPRA Thesaurus* , ya que la institución desarrollo una base de datos con una gran

cantidad de información sobre revistas especializadas en el campo de las ciencias de los polímeros, *Polymer Library* y se necesitaba un vocabulario acorde a la gestión de toda esa gran cantidad de información .

Polymer ontology u ontología de polímeros

Cualquier ontología que se desarrolle en el campo de los polímeros debe tener en cuenta todas estas facetas posibles en el desarrollo de una jerarquía y de la estructura de los diferentes conceptos. Tomamos como ejemplo la Polymer Ontology, desarrollada por *Anna-Maria Schoeller*, que contempla el siguiente esquema de clasificación:

1-Manufactura de procesos:

- *Polimerización:* conversión de monómeros en polímeros lineales, ramificados o cíclicos
- *Policondensación:* reacción química de monómeros para formar polímeros por medio de reacciones de condensación (separación de productos de bajo peso molecular)
- *Copolimerización o poliadición:* los monómeros son añadidos a la estructura básica, a diferencia de la policondensacion

2- Orígenes de las materias primas:

- *Materiales naturales modificados*
- *Materiales sintéticos*

3- Composición de las macromoléculas

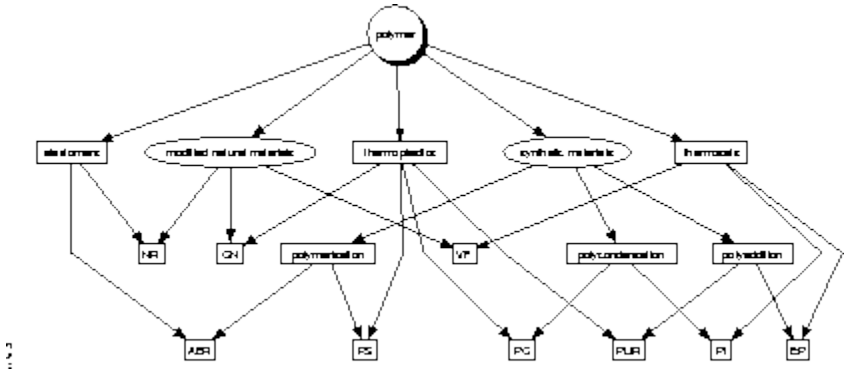
- *Estructura amorfa*
- *Estructura semicristalina*
- *Estructura liquida cristalina*

4. Tecnología de procesamiento y propiedades de polímeros

- Termoplásticos: materiales con conexiones débiles y que al ser calentados pueden reutilizarse

- Termoestables: materiales con enlaces cruzados que luego de calentarse no pueden ser reutilizados (curado)

- Elastómeros: materiales con orientación irregular en las cadenas atómicas que luego de una deformación inicial retornan a su forma original



Polyinfo RDF

En el caso de la base de datos de polímeros Polyinfo, se utilizó el formato RDF para crear descripciones semánticas de las reacciones de formación de polímeros o polimerización, creando una correlación con los monómeros y un enlace conceptual con sustancias relacionadas en otras bases de datos como Nikkaji (Japan Chemical Substance Dictionary), la cual es fundamentalmente una base de datos con información sobre monómeros de aproximadamente 3,459,747 registros En este caso no se utilizó una ontología pero se efectuó una correlación entre los monómeros y las reacciones de polimerización correspondientes entre las dos bases de datos con una probabilidad de 94%.

El enlace entre PolyinfoRDF y Nikkaji RDF se realizó mediante especificaciones en SKOS utilizando las etiquetas de equivalencia cercana *skos:closeMatch* entre conceptos de las bases de datos, teniendo en cuenta los monómeros, las

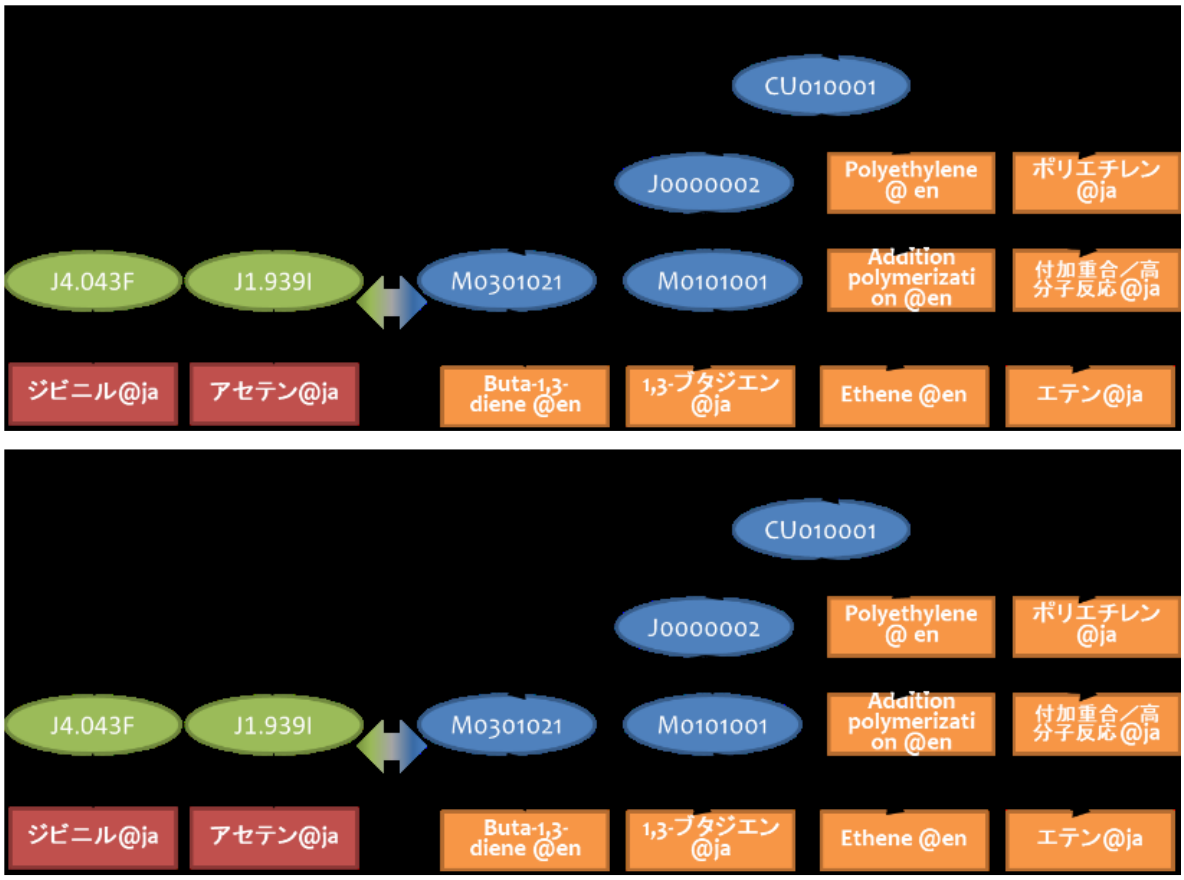
reacciones de polimerización y los polímeros correspondientes vinculando los ID o números de identificación de cada base en los grafos. En este caso, el concepto más importante para establecer un enlace con otras fuentes es el nombre del monómero, ya que: los polímeros se sintetizan fundamentalmente a partir de sus monómeros constituyentes y la síntesis o reacción de polimerización por otro, representa la información fundamental para la obtención del polímero y existían previamente diferentes bases de datos con nombres de monómeros relacionados con el campo de ciencias biológicas en RDF.

La base utiliza números originales de identificación (IDs) para los nombres de los polímeros, reacciones de polimerización y nombres de monómeros y con ellos se pudieron diseñar protocolos para el establecimiento de grafos en RDF.

Por ejemplo, en el caso del poliestireno identificado por el *ID CU010001, (CU010001 ns2:label "Polyethylene"@en.)*. A su vez, se le asigna una reacción de polimerización con *ID J000002 (CU01000ns1:pHasPolymerizationPath J000002.)* A la que a su vez se le asigna la etiqueta de polimerización de adición *(J000002 ns2:label "Addition polymerization"@en.)*, asociado con el monómero estireno con ID *M0301021 (J000002 ns1:pHasMonomer M0301021, M0101001* y así sucesivamente, vinculando conceptos con propiedades en RDF.

En el caso de, por ejemplo, el butandieno con ID M0301021 y el eteno M0101001, los equivalentes cercanos en la base Nikkaji son J4.043F y J1.939I, respectivamente y los tripletes RDF en SKOS se expresan:

@prefix skos: <http://www.w3.org/2004/02/skos/core#>
ns1: M0301021 skos:closeMatch nikkaji: J4.043F
ns1: M0101001 skos:closeMatch nikkaji: J1.939I



Pero al no ser entidades químicamente idénticas en este caso la etiqueta de equivalencia aproximada es una mejor aproximación que la etiqueta de equivalencia en SKOS y de OWL, "*skos:exactMatch*" "*owl:sameAs*".

No obstante, el éxito alcanzado, se detectaron problemas de incompatibilidad con los nombres de los monómeros según la nomenclatura de IUPAC (Unión Internacional de Química Pura y Aplicada) y la sintaxis usada en RDF, que incluyen frecuentemente las dobles comillas y ello puede crear un error de sintaxis en RDF.

Pasos para la implementación de un vocabulario controlado en open linked data

La aplicación de la tecnología de datos enlazados a los lenguajes documentales permite mejoras en la recuperación de información en varios idiomas, realizar búsquedas temáticas especializadas facetadas, navegación transversal entre disciplinas y campos relacionados, etc. También una reutilización de estos y una extensión de vocabularios más generales a otros más específicos.

En primer lugar, la primera cuestión a analizar es si utilizar un lenguaje o vocabulario preexistente o desarrollar uno nuevo. Dada la complejidad de la segunda opción, la recomendación apuntaría hacia la primera opción.

La segunda cuestión a determinar en la transformación de un vocabulario controlado en un formato apropiado a los datos enlazados es si usar SKOS u OWL, la recomendación también seria por la primera opción por su menor complejidad aun teniendo en cuenta sus limitaciones, además en el caso de los tesauros SKOS esta perfectamente adaptado para la representación de todas sus relaciones semánticas y permite añadir más relaciones no originalmente contempladas en el desarrollo del mismo y se ha convertido en una norma específicamente diseñada para la adaptación y desarrollo de este tipo de vocabularios, siendo más adecuado para un modelado semántico de los mismos. SKOS permite modelar adecuadamente las relaciones de equivalencia, las relaciones jerárquicas, tanto las de género especie como las partitivas, y las relaciones asociativas, aunque se adapta a tesauros y encabezamientos de materias.

Pero debe recordarse que SKOS permite una separación de dominios o de espacios de representación, es decir, por un lado se establece la noción de concepto según la norma ISO 25964 y por otro existe una separación estructural de otros ámbitos de aplicación como *skos:Collection*, *skos:ConceptScheme* , es decir, la estructura general le asigna a cada concepto una posición especifica dentro de un esquema general (*skos:inScheme*) como parte integrante de la estructura conceptual de un esquema de clasificación, puede ser el concepto

más genérico de una cadena jerárquica *skos:hasTopConcept* o como parte integrante de la jerarquía *skos:Concept*. Pero en la colección definida por el modelo *skos:Collection*, no permite la agrupación en el mismo nivel o establecer relaciones semánticas con otras estructuras, pero si agrupar conceptos en un campo semántico especifico, lo cual es muy importante para la aplicación en dominios temáticos más específicos, en microtesauros o en subdivisiones de materias más especializadas como es el campo de las Ciencias de los Polímeros, por lo que es ideal para la representación de vocabularios controlados en dominios temáticos específicos.

En el caso de una clasificación especializada puede tenerse en cuenta la alternativa de usar OWL, aunque SKOS permite la generación de esquemas generales (etiqueta *ConceptScheme*) y permite además la representación de todas las relaciones semánticas que pueden darse en una clasificación a (relaciones poli jerárquicas, notas de alcance o de definición, agrupaciones semánticas de conceptos, etc.)

Una vez seleccionado el lenguaje de representación para el vocabulario, debe evaluarse la posibilidad de la utilización de distintos programas de gestión de vocabularios controlados, de acceso libre o pagos, como PoolParty, TemaTres, etc. En distintos aspectos o parámetros (funcionalidad, capacidad semántica de representación, actualización, facilidades de importación y exportación, capacidades de publicación, etc.)

En una tercera instancia, deben considerarse el tipo de recursos con los que se debe establecer un enlace al vocabulario, sobre todo recursos considerados de autoridad con una gran cantidad de información sobre el tema específico cubierto por el vocabulario, pero debe tenerse en cuenta el problema de la granularidad alcanzada por el lenguaje documental en temas específicos y la disparidad que puede encontrarse en distintas fuentes de información, que pueden no tener el nivel de granularidad desarrollado por el vocabulario.

Conclusiones

El uso de linked data permite utilizar recursos web para aumentar el conocimiento disponible sobre un concepto o tema, permite ofrecer links o enlaces a fuentes externas de una gran cantidad de información como DbPedia o Wikidata, artículos académicos, definiciones, formatos multimedia, etc. Y pueden ser añadidas como un recurso extra al portal de la propia organización, permite añadir links y recursos a organizaciones relacionadas con la temática de la propia institución y a su vez permitir referenciar a otras instituciones a la misma y añadir conjuntos complejos de datos en sistemas de conocimiento especializados a la comunidad que comparte datos enlazados.

No obstante, debe tenerse en cuenta que fuentes de autoridad en materia de datos enlazados como Dbpedia a menudo no tendrán el control de la sinonimia, o el nivel de especificidad requerido en las relaciones jerárquicas representadas por un vocabulario especializado en un dominio científico especifico, como es la ciencia de los polímeros, que puede desarrollar un sistema de clasificación especializado o un tesauro, por lo que una primera aproximación sería establecer en el caso de un tesauro por ejemplo, en el nivel del mismo a una fuente externa, como Dbpedia, es decir, frente a un cuerpo de conocimientos especializado, en este caso Ciencia de los Polímeros y contando con vocabularios controlados especializados en el campo, sean esquemas de clasificación o tesauros, el resultado ideal es que cada concepto de la estructura conceptual del vocabulario corresponda a una URI de una fuente externa, como por ejemplo Dbpedia, sobre el mismo tema, por ejemplo. Si se toma el término del tesauro *Plásticos*, un término jerárquicamente dependiente del término general *Polímeros*.

El primer paso sería crear en el software gestor del vocabulario que se utilice dentro del término un campo que incluya la URI de otras fuentes externas como DbPedia por ejemplo http://dbpedia.org/page/Plastic, lo que permitiría añadir

información automáticamente definiciones, imágenes, noticias, actualizaciones o enlaces hacia otras publicaciones sobre la temática, a través del uso de SPARQL endpoints, para hacer búsquedas en ese contexto o para referenciar o enlazar paginas dentro de la Dbpedia. Este proceso manual, puede ser automatizado por medio de spotlights (términos con alto grado de confianza que tienen un recurso asociado en Dbpedia, con una URI correspondiente) y ser validado posteriormente, lo que se realiza por la aproximación semántica más que en la coincidencia exacta.

No obstante, el nivel de granularidad posible en este tipo de fuente puede ser mucho más limitado de lo que se requiere para un campo especializado. Si tomamos el termino *Polietileno*, por ejemplo, si podemos encontrar una entrada correspondiente, http://dbpedia.org/page/Polyethylene, pero no es posible seguir a un nivel de granularidad mayor, síntesis, copolímeros, etc., por lo que constituye una limitación a tener en cuenta.

En el caso de que el nivel de granularidad del vocabulario sea mayor que los conceptos encontrados en la fuente referenciada, una posible solución sería incluir el término más especifico en su correspondiente termino más general, aunque en ese nivel la correspondencia se daría en un nivel más general entre recursos enlazados y términos generales de la jerarquía,

En las fases a considerar para el desarrollo o la adaptación de un vocabulario a formatos de open linked data, entonces se deben tener en cuenta aspectos como el formato a usar para la representación compatible con el marcado semántico, la vinculación y el acceso a fuentes de información de reconocida autoridad sobre la misma temática, a bases de datos especializadas en la misma, repositorios digitales, archivos, otras bibliotecas y un gestor adecuado del vocabulario controlado, de licencia libre. Permitir el agregado manual pero fundamentalmente la vinculación automática a conjuntos de datos, lo que permitiría el descubrimiento de nuevo conocimiento y nuevos recursos, lo que

ofrece nuevas posibilidades al acceso al conocimiento en una disciplina especializada, sobre todo cuando se vinculan fuentes y conjuntos de datos de calidad reconocida.

Para la inclusión de recursos deben tenerse en cuenta la opinión de especialistas y usuarios de las fuentes y recursos de información de un campo disciplinario, deben tenerse en cuenta bases de datos y recursos especializados en los polímeros y materiales relacionados de acceso libre como por ejemplo, MATWEB (On Line Material Resources, http://www.matweb.com/), OMENXUS (Free On Line Database for Plastic Industry, https://omnexus.specialchem.com/), entre otras, de forma de poder establecer vínculos entre los monómeros, sus propiedades correspondientes y los polímeros resultantes, así como también las reacciones de síntesis que permiten obtenerlos a partir de sus monómeros correspondientes, ya que el modelo de datos enlazados es un enfoque para la codificación de datos con un alto grado de granularidad, requisito que no satisfacen las principales fuentes de información general disponibles en la web, como en el caso de Dbpedia, para un campo especializado.

También en una segunda etapa puede considerase el establecimiento de concordancias o mapeos entre distintos sistemas de organización del conocimiento del sector, taxonomías, tesauros, ontologías, para aumentar la capacidad de recuperación de recursos y su disponibilidad si son de acceso libre, sobre todo, teniendo en cuenta que pueden establecerse relaciones entre entidades en diferentes esquemas por las equivalencias entre etiquetas de SKOS, OWL y RDF otros formatos de codificación semántica, lo que permite la alineación entre diferentes vocabularios.

Finalmente, el modelo de datos enlazados permite no solo implementar la publicación de vocabularios especializados en las webs legibles por personas sino además para su lectura por agentes de software, para poder permitir el

descubrimiento automático de nuevas relaciones y recursos de información, creando de esta forma una nube de datos enlazados en distintas especialidades.

La dificultades en el modelo estarían en que al ser un campo especializado, las bases de datos y fuentes especializadas en la temática no son de acceso libre sino pago y en la actualización de los enlaces establecidos, no obstante, la aplicación del formato de datos enlazados a los vocabularios controlados permiten un enriquecimiento y una mejora notable en la capacidad de recuperación de información en distintos idiomas y en una mayor cantidad de fuentes de otras disciplinas relacionadas, su reutilización en otros contextos más específicos y servicios de alineamiento para la gestión de actualizaciones que logran versiones estables en tiempo real.

Además debe recordarse que si bien en general las bibliotecas en general no preparan o consideran que los datos bibliográficos y de categorización semántica no son necesarios al menos con un gran nivel de granularidad, las necesidades de usuarios especializados en un campo científico muy especializado, como es el campo de la Ciencia de los Polímeros, si requieren muy a menudos metadatos y una codificación semántica con un mayor grado de granularidad que campos más generales o amplios de una disciplina científica determinada, por lo que el aplicar esta tecnología a vocabularios especializados de esta especialidad puede resultar en una mejora considerable de los mismos para la recuperación de información especializada.

Bibliografía

Ávila Alonso, Rafael. *Aplicación de los principios Linked Open Data a la lista de encabezamientos de materia de Biblioteca de la Universidad Politécnica de Madrid.* Madrid: Universidad Carlos III, sep,2014, Disponible en: https://earchivo.uc3m.es/bitstream/handle/10016/19674/avila_aplicacion_TFM_2014.pdf?sequence=1&isAllowed=y, (Acceso 2 de mayo 2020)

Blaney, Jonathan. *Introduccion a los Datos Abiertos Enlazados*. Disponible en : https://programminghistorian.org/es/lecciones/introduccion-datos-abiertos-enlazados, (Acceso 2 de mayo 2020)

Casanovas, Ines. *Gestión de Archivos Electrónicos*. Buenos Aires: Alfagrama, 2008

Ishii, Masashi, Taro Takemura y Mikiko Tanifuji. *PoLyInfo RDF: A Semantically Reinforced Polymer Database for Materials Informatics*. Disponible en: http://ceur-ws.org/Vol-2456/paper18.pdf, (Acceso 2 de mayo 2018)

Méndez, Eva y Jane Greenberg. Datos Enlazados para Vocabularios Abiertos y Marco General de HIVE. *El profesional de la Información*, 2012, 21(3); 236-244, Disponible en: http://www.elprofesionaldelainformacion.com/contenidos/2012/mayo/03_esp.pdf, (Acceso 2 de mayo 2020)

Peña Vera, Tania. *Organización y Representacion del Conocimiento.* Buenos Aires: Alfagrama, 2011

¿Qué es SKOS? Disponible en: https://skos.um.es/acerca/index.php, (Acceso 2 de mayo 2020)

RAPRA Classification Code. En: BARTOC, Disponible en: https://www.bartoc.org/sv/node/1084, (Acceso 2 de mayo 2020)

RAPRA Thesaurus. Rubber and Plastics Research Association of Great Britain, 1985

Smith, Melody. Smart Thesauri: Using Taxonomies with Linked Data En: *Taxodairy*, jul.2015, Disponible en: https://taxodiary.com/2015/07/smart-thesauri-using-taxonomies-with-linked-data/ (Acceso 2 de mayo 2020)

Tesauro de Materias de SERBIULA. Disponible en: http://www.serbi.ula.ve/tematres/vocab/index.php, (Acceso 2 de mayo 2020)

The Ontology of Polymers, Disponible en: http://www.dfki.uni-kl.de/~imcod/htdocs/Bernd/Paper/paper/node5.html, (Acceso 2 de mayo 2020)