Profesorado
Revista de currículum y formación del profesorado

VOL. 15, Nº 1 (2011)
ISSN 1138-414X (edición papel)
ISSN 1989-639X (edición electrónica)
Fecha de recepción 19/08/2010
Fecha de aceptación 22/10/2010

INTEGRANDO LAS TECNOLOGÍAS DE WEB SEMÁNTICA EN LA ARCHIVÍSTICA

Integrating Semantic Web Technologies in Archival Science

José Manuel Morales-del-Castillo y Germán Hurtado Martín*
** Universidad Carlos III de Madrid*
*** Universidad de Gante (Bélgica)*
E-mail: jmdel@bib.uc3m.es, german.hurtadomartin@ugent.be

Resumen:
La Archivística tradicional se enfrenta al reto de adaptar sus principios teóricos y heurísticos a entornos digitales de trabajo. La adopción de nuevas tecnologías de la información, como las tecnologías de Web Semántica, pueden abrir la puerta a la mejora de la descripción, acceso y recuperación de información en este tipo de plataformas. En este trabajo se propone la utilización de vocabularios semánticos y metodologías para la creación y mantenimiento de ontologías web aplicadas al desarrollo de herramientas archivísticas. Por último, se concluye que la incorporación de estas tecnologías y metodologías al currículo de los archiveros se puede hacer de una manera natural, complementando competencias y habilidades que ya adquieren estos profesionales durante su formación reglada.
Palabras clave: Archivística; Web Semántica; Ontologías.
Abstract:
Traditional archival science has to face the challenge of adapting its theoretical and heuristic principles to be applied in digital-based work environments. The adoption of new information technologies, as Semantic Web technologies, can provide the chance to improve information description, access and retrieval in these platforms. In this paper using authors propose apply several semantic languages and methodologies to design and manage web ontologies applied to develop archival tools. Finally, it's concluded that incorporating these technologies and methodologies into archivist's curriculum can be done in a natural way, thus complementing competences and skills that these professionals already acquired along their academic training.
Key words: Archival Science; Semantic Web; Ontologies.

Agradecimientos: Este trabajo ha sido financiado dentro del Programa de Ayudas a la Movilidad del Plan Propio de Investigación de la Universidad Carlos III de Madrid.

1. Introducción

Tradicionalmente la enseñanza de la Archivística se ha visto lastrada por un cierto inmovilismo plasmado en los planes de estudio de los centros universitarios españoles donde se imparten estudios de Biblioteconomía y Documentación, y abocada en numerosas ocasiones a repetir el mismo discurso que la lleva a moverse entre el enfoque tradicional y la preservación de la documentación histórica, y el recelo que despierta en muchos archiveros la gestión de registros usando nuevas tecnologías en un ámbito que históricamente nunca antes había necesitado de ellas. Es muy común oír hablar de cómo se intenta ofrecer una visión integral e integrada de estas dos corrientes en una sola disciplina, pero muchas veces casi se hace más hincapié en remarcar las diferencias que existen entre ellas que en poner de relieve los puntos que tienen en común. No obstante, esto no nos debe resultar del todo extraño ya que hasta ahora para la Archivística tradicional sólo existía un marco teórico y heurístico bien definido, pero no una metodología, unas herramientas o unos lenguajes que realmente puedan ofrecer una solución que concilie ambas tradiciones archivísticas.

Esta situación de cierto anquilosamiento puede verse agravada ante el nuevo reto que supone la inexorable implantación de las tecnologías de la información en todos los ámbitos de la sociedad y que, entre otros efectos, está provocando que los sistemas tradicionales de información se estén moviendo hacia nuevos entornos y plataformas de trabajo. Debido a esta tendencia, es cada vez más frecuente ver cómo algunas bibliotecas han dado ya el paso hacia su transformación en bibliotecas digitales, y cómo determinadas administraciones públicas están adaptando su forma de dar servicio a los ciudadanos dentro del marco definido por el amplio concepto *e-gobierno* (Palvia y Sushil 2007). En este escenario, donde las barreras físicas se difuminan, se hace imprescindible trabajar con nuevos estándares para representar y compartir la información, y desarrollar mecanismos eficaces para la gestión, reutilización y recuperación de recursos que provienen de fuentes heterogéneas.

Un ejemplo de tecnologías que podrían servir como base para el diseño de herramientas con estas funcionalidades son los lenguajes basados en XML (eXtensible Markup Language) (Bray et al. 2008), que se fundamentan en el marcado de recursos mediante etiquetas, y que abren la puerta a la posibilidad de representar como *conocimiento* tanto los procesos que se desarrollan en una organización (ya sean administrativos o de negocio), como los diferentes agentes o artefactos documentales que están asociados a estos. De esta manera es posible optimizar el acceso a la información y articular mecanismos de gestión más precisos.

Esta filosofía de representación de la información la asume el proyecto Web Semántica (Berners-Lee, Hendler y Lassila 2001), cuyo objetivo consiste en extender el actual modelo de Web utilizando una serie de vocabularios que permiten enriquecer la descripción de los recursos disponibles en la red y de esta forma hacerlos semánticamente accesibles. De esta manera, la Web se convierte en una plataforma universal para el intercambio de información. Para conseguirlo, el proyecto se basa sobre dos pilares básicos:

- El etiquetado semántico de recursos de manera que la información pueda ser interpretada tanto por humanos como por máquinas.

- El desarrollo de agentes software inteligentes (Hendler 2001) capaces de operar a nivel semántico con los recursos web e inferir nuevo conocimiento a partir de ellos (pasando de

esta manera de la búsqueda de palabras clave a la recuperación de información mediante conceptos). Los agentes pueden procesar estos recursos utilizando unas herramientas denominadas ontologías (que describiremos más adelante) que permiten desambiguar su semántica al contextualizarlos en un dominio concreto.

A la vista de las capacidades que ofrece el desarrollo del modelo de Web Semántica, la aplicación de las tecnologías sobre las que se basa puede ser de gran utilidad en un área como el de los archivos donde la preservación del contexto de producción de la documentación generada es un elemento clave para la descripción, organización y gestión de los fondos. Un ejemplo de la aplicación de las tecnologías de Web Semántica a los archivos lo encontramos en el trabajo de Palacios, Cremades y Costilla (2005) donde se integran las ontologías en el funcionamiento de un archivo parlamentario. Estos autores describen un sistema de gestión de archivos que se basa en la definición de ontologías que permiten controlar la descripción de documentos (utilizando metadatos Dublin Core (2010), y las normas de descripción archivística ISAD (ICA 2000) e ISAAR (ICA 2004)), y ontologías específicas del dominio de trabajo del archivo para conseguir una gestión más eficaz de la documentación.

No obstante, el uso de herramientas para capturar el contexto no es algo nuevo en los archivos. Sin ir más lejos, los cuadros de clasificación son elementos fundamentales en la Archivística moderna que permiten estructurar lógicamente el fondo a partir del análisis funcional de la actividades que desarrolla la institución. Entonces, ¿por qué no aplicar tecnologías semánticas para el desarrollo de herramientas archivísticas de esta naturaleza?

En este trabajo pretendemos dar respuesta a esta cuestión presentando una metodología que engloba diversos procedimientos, herramientas y lenguajes que permiten la integración de las tecnologías semánticas en los archivos. Además intentamos poner en valor la figura del archivero, como conocedor que es de la estructura lógica del fondo con el que trabaja, como un activo que hay que aprovechar para desarrollar archivos más eficientes y eficaces.

El resto del trabajo se estructura de acuerdo al siguiente esquema. En el apartado 2 presentamos algunos conceptos archivísticos que sirven como punto de partida para introducir la transición del cuadro de clasificación a las ontologías, que comentamos en el apartado 3. En el apartado 4 esbozamos cómo generar ontologías con un vocabulario específico. En los apartados 5 y 6, respectivamente, comentamos dos metodologías que se pueden aplicar para refinar el diseño de ontologías y mantenerlas en un sistema de información. Por último, en el apartado 7 apuntamos algunas conclusiones.

2. Conceptos básicos de archivística

En primer lugar es importante tener claros algunos conceptos archivísticos básicos que están relacionados con la forma en que se define la estructura lógica de un fondo y que nos ayudarán a situar nuestra propuesta en el marco teórico adecuado.

Según define Roberge (1993) en un archivo podemos distinguir las siguientes divisiones y subdivisiones que permiten la organización lógica de la documentación de una institución:

- Fondo: Se define por el principio de procedencia y está compuesto por la totalidad de la documentación producida o recibida por la institución en el tiempo. A su vez, es

posible diferenciar subdivisiones del fondo (sub-fondos) que deberán ser tratados, clasificados, ordenados y descritos de manera independiente.

- Sección: Es la división primaria del fondo de acuerdo a las funciones, actividades o líneas de actuación generales que lleva a cabo la entidad.

- Subsección: Es la división de una sección realizada en virtud de las actividades que se desarrollan dentro de las funciones genéricas de la institución. A su vez pueden subdividirse cuando puedan distinguirse otras actividades independientes más específicas.

- Series: De una manera formal, según Cruz Mundet (2005), las series documentales se pueden definir como un conjunto de unidades archivísticas (expedientes, libros, etc.) agrupadas por ser el resultado de una misma actividad, y que han sido producidas y agrupadas de manera continua (seriada) en el proceso de tramitación administrativa.

Basándose en estos elementos (con los que es posible establecer el origen funcional de los documentos generados o recibidos por una institución) y en una serie de principios teóricos, como el principio de respeto al orden original y el principio de procedencia (cuyo objetivo es mantener y conservar el contexto en el cual se produce la documentación), en el ámbito de la Archivística se han desarrollado herramientas, como los cuadros de clasificación, que facilitan la localización y acceso a los documentos del fondo. Los cuadros de clasificación permiten definir la organización jerárquica y lógica del fondo de un archivo a partir de las funciones o actividades que se realizan en la institución y fruto de las cuales se crean o reciben documentos. El cuadro de clasificación permite organizar intelectualmente la información y situar los documentos en su contexto de producción mediante la definición de las relaciones que estos establecen entre sí.

Para poder elaborar esta herramienta (una por cada sub-fondo que componga el fondo del archivo) es necesario que previamente se conozcan las funciones del organismo que genera la documentación, o lo que es lo mismo, conocer su contexto de creación. Para facilitar esta tarea, las funciones, actividades o transacciones que se desarrollan en la institución se harán corresponder con las secciones (y subsecuentes subdivisiones) que aparecerán en el cuadro de clasificación y que dan lugar a las series documentales, quedando el origen funcional trazado de una manera similar a la siguiente:

- *Sección* (función): Organización y administración

- *Sub-sección* (actividad): Administración general

- *Serie:* Memorias

- *Sub-serie:* Memorias de investigación

Cada una de las series definidas en el cuadro posee identidad propia y se relacionará jerárquicamente con las demás de forma que no quepa ambigüedad posible a la hora de identificarlas (Ruiz 1995). Procediendo de esta manera, y definiendo el origen funcional de todas y cada una de las series que se pueden distinguir en un fondo, obtendríamos como resultado el cuadro de clasificación para ese fondo específico.

3. La transición del cuadro de clasificación a la ontología

Como vemos, los cuadros de clasificación son básicos en el funcionamiento de los archivos tradicionales, por lo que sería de vital importancia garantizar en el entorno web la existencia de herramientas que cumplan las mismas (o similares) funciones. En el área de la inteligencia artificial encontramos los sistemas de organización de conocimiento que permiten representar parcelas de realidad y desarrollar con la información operaciones de razonamiento más o menos complejas. Este es en esencia el objetivo que cumplen los cuadros de clasificación en un archivo, ya que con ellos podemos definir el contexto de creación de los documentos para gestionarlos, localizarlos y acceder a ellos de una manera más eficaz. Uno de los sistemas de organización de conocimiento más utilizados en el desarrollo de sistemas de información que se apoyan sobre bases de conocimiento son las denominadas ontologías, las cuales, tal y como hemos comentado anteriormente, se pueden considerar como uno los pilares básicos sobre los que se apoya el proyecto Web Semántica.

Según la RAE, en su origen el término "*ontología*" denomina una rama de la filosofía que se encarga del estudio del ser en general, de sus propiedades trascendentales y, más específicamente, de la organización de la realidad. Como derivación de esta última acepción encontramos la definición de las ontologías entendidas como sistemas de organización del conocimiento. Según diferentes autores (Neches 1991) (Gruber 1995) (Guarino 1998) las ontologías pueden ser entendidas como la suma de una serie de conceptos relevantes que representan el conocimiento compartido por los miembros de un dominio concreto, las relaciones que establecen entre sí estos conceptos, y los axiomas definidos sobre estos conceptos y relaciones. Las ontologías suponen en sí un canal de comunicación entre personas y máquinas (Clyde 2002) ya que permiten establecer un puente entre el lenguaje natural y la manera en la que se comunican entre sí las máquinas (Jenz 2003).

Según Gruber (1995), las ontologías web presentan una serie de componentes estructurales básicos:

- Conceptos: Son las ideas básicas que se intentan formalizar. Los conceptos pueden ser clases de objetos, métodos, planes, estrategias, procesos de razonamiento, etc.

- Relaciones: Representan la interacción y enlaces que se define entre los conceptos de un dominio. Por ejemplo: *subclase-de*, *parte-de*, *parte-exhaustiva-de, conectado-a*, etc.

- Instancias: Se utilizan para representar objetos o individuos determinados pertenecientes a un concepto (o clase). Por ejemplo, *Mario Benedetti* sería una instancia de la clase *Poetas latinoamericanos*. Como resultado de la instanciación de una ontología obtenemos la base de conocimiento sobre la que el sistema puede operar y realizar inferencias.

- Axiomas: Son asertos que se declaran sobre relaciones que deben cumplir los elementos de la ontología. Por ejemplo: "*Si A y B son de la clase C, entonces A no es subclase de B*", etc.

Muchos de estos elementos no son en absoluto ajenos a los profesionales titulados en las facultades de Biblioteconomía y Documentación españolas, ya que como parte de su formación han recibido nociones sobre lenguajes documentales controlados como los tesauros, que definen una serie de relaciones semánticas de carácter jerárquico, asociativo y de equivalencia entre los conceptos más relevantes de un dominio específico (Arana y Codina

2004). Desde este punto de vista, los tesauros no dejan de ser un esquema de conocimiento similar a las ontologías, aunque con menor capacidad descriptiva.

4. Cómo definir ontologías en la Web

Existen multitud de lenguajes que permiten representar ontologías como CASL (*Common Algebraic Specification Language*) desarrollado por la Universidad de Bremen (2008) o CycL (Cycorp 2002) pero la mayoría de ellos no están específicamente diseñados para ser usados en la Web. Para este contexto específico el estándar recomendado por el W3C (World Wide Web Consortium 2010) es OWL (MacGuinnes y van Harmelen 2004), un lenguaje orientado al desarrollo de ontologías web y cuyas siglas, aunque con el orden alterado, se corresponden con *Web Ontology Language.* En esencia es un lenguaje que permite describir semánticas procesables de forma automatizada utilizando motores de inferencia sin imponer restricciones sobre el razonamiento que se puede realizar con los recursos, garantizando de esta forma su interoperabilidad (es decir, los recursos pueden ser reutilizados en diferentes procesos y aplicaciones).

El lenguaje se construye sobre el modelo de datos que proporciona el vocabulario RDF (*Resource Description Framework*) (Becket 2004) y sobre las proposiciones lógicas de RDF Schema (Brickley y Guha 2004), un vocabulario que permite definir la semántica de los asertos definidos en RDF. En otras palabras, OWL extiende la semántica que es posible definir con RDF Schema y adopta la manera que tiene RDF para estructurar la información para describir en la Web recursos, objetos físicos, conceptos abstractos o cualquier otra cosa con identidad propia.

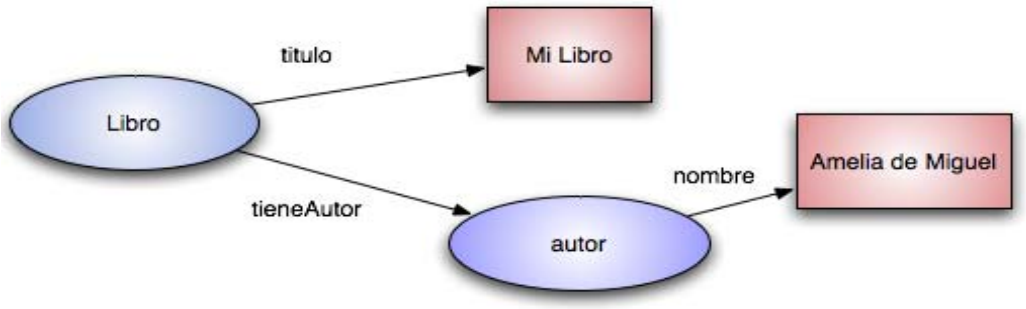
La información se representa mediante sentencias que toman la forma de tripletas *sujeto-propiedad-objeto*, donde el sujeto indica a *qué* o *quién* se refiere la sentencia (es decir, el concepto o recurso que queremos describir), la propiedad indica una característica determinada del sujeto, y el objeto indica el valor que toma la propiedad (que puede ser un literal o un nuevo concepto). Por lo tanto, se pueden distinguir dos tipos de tripletas atendiendo a los elementos que las conforman: *recurso-propiedad-recurso* y *recurso-propiedad-valor*. Esta descripción se realiza en función de las propiedades que caracterizan a los recursos y los valores que estas toman, de forma que es posible enriquecer la definición de un recurso simplemente añadiéndole nuevas propiedades.

Aunque existen varias formas de representar las proposiciones RDF la sintaxis más extendida es la basada en el metalenguaje XML (eXtensible Markup Language) (Bray et al. 2008), que en los últimos años se ha convertido en un estándar para el intercambio de información en la Web, y que ya está presente en muchos planes de estudios de las facultades españolas que imparten la titulación de Biblioteconomía y Documentación. Ese es el caso, por ejemplo, de la asignatura optativa "Archivos electrónicos" que se imparte en la Universidad de Granada y en la que XML se utiliza para representar registros archivísticos utilizando la norma EAD (*Encoded Archival Description*) de la Library of Congress (2002).

No obstante, las tripletas RDF pueden también representarse de manera gráfica utilizando grafos orientados de acuerdo a la siguiente convención (ver figura 1):

- los conceptos se representan mediante elipses,

- las propiedades como líneas que conectan conceptos con otros conceptos o con literales, los literales se representan como rectángulos.



- **Figura 1.** *Ejemplo de grafo orientado*

Además, OWL define semánticas y reglas básicas de razonamiento basadas en la lógica de predicados y la lógica descriptiva entre las que encontramos las siguientes (Antinou, van Harmelen 2004):

1. *Pertenencia a clases*: Si *x* es una instancia de la clase C, y C es una subclase de D, entonces *x* es una instancia de la clase D.

2. *Equivalencia de clases*: Por ejemplo, si A es una clase equivalente a la clase B, y B es equivalente a la clase C, entonces A es equivalente a C.

3. *Consistencia del conocimiento*: Permite determinar errores lógicos en la definición de clases e instancias. Por ejemplo, si declaramos una instancia *x* que pertenece a la clase A, y A es a su vez subclase simultáneamente de la clase B∩C y de la clase D (siendo D y B clases disjuntas), entramos en contradicción ya que A debería ser una clase vacía.

4. *Clasificación de instancias*: La imposición de determinados pares *propiedad/valor* como condición suficiente para ser miembro de una clase, nos permite definir tantas instancias como sujetos cumplan esa condición.

Sin embargo, y a pesar de sus muchas capacidades, en OWL no es posible definir reglas de inferencia complejas por lo que en ese caso deberíamos recurrir a lenguajes de reglas específicos que permiten integrarlas dentro del código de la ontología, como es el caso, por ejemplo, de SWRL (Horrocks et al. 2004). Por esta razón, en este trabajo nos vamos a centrar exclusivamente en la tarea de generar la estructura semántica del sistema, dejando de lado el proceso de definición de reglas, que pensamos queda fuera del alcance del mismo.

Por otro lado, usar un lenguaje con una gran capacidad expresividad como OWL tiene también como contrapartida la imposibilidad de implementar las ontologías de una manera eficiente si estas se diseñan sin definir ningún tipo de restricción sobre los constructores o instrucciones de OWL que se pueden utilizar. Para conciliar capacidad expresiva con eficiencia de procesamiento el grupo de trabajo del *World Wide Web Consortium* que se encarga del desarrollo normativo de OWL se vio en la necesidad de definir tres sub-lenguajes

(o especies) para OWL con el fin de poder ofrecer a los desarrolladores diferentes combinaciones de *expresividad-capacidad de razonamiento* que permitieran adaptar las ontologías a la complejidad y necesidades específicas de un sistema determinado:

- OWL Full es la especie de mayor capacidad expresiva ya que integra todas las primitivas definidas en el lenguaje (para ser rigurosos ni siquiera debería considerarse como un sub-lenguaje) y permite combinarlas de forma arbitraria con RDF y RDF Schema.

- OWL DL (*Description Logics*) por su parte restringe el modo en que se pueden utilizar los constructores de OWL y RDF de manera que las ontologías se ajustan a los requerimientos de los motores de inferencia basados en lógica descriptiva.

- OWL Lite impone condiciones más estrictas sobre los elementos de OWL que se pueden utilizar, lo cual disminuye su capacidad expresiva pero facilita el razonamiento con las ontologías diseñadas usando este sub-lenguaje.

Como vemos, la elección de una u otra especie depende del problema que queramos resolver y de cuáles son las necesidades específicas del sistema en el que queremos implantar nuestra ontología (o lo que es lo mismo, hay que tener claro si la prioridad de nuestro sistema es la interoperabilidad o la facilidad de procesamiento). Por esta razón, a día de hoy las implementaciones prácticas de ontologías se hacen utilizando OWL DL u OWL Lite, ya que con OWL Full son inviables.

4.1. Elementos y características de OWL

Una vez que conocemos las principales características de OWL, vamos a hacer a continuación un repaso de los conceptos básicos que nos permiten definir la estructura de una ontología usando este vocabulario.

4.1.1. Clases

Dado que en RDF la información se estructura en tripletas *sujeto-propiedad-objeto* podríamos, por ejemplo, definir el aserto *Juan tiene como primer apellido López*. No obstante, si queremos generalizar este conocimiento y expresar, por ejemplo, que existen otros seres (sin referirnos a ninguno en concreto) que también tienen *primer apellido,* entonces necesitamos recurrir al concepto de clase.

Una clase se puede definir como un ente lógico que aglomera un conjunto de individuos que presentan una serie de propiedades comunes, y donde cada uno de esos individuos sería una instancia de la clase. Así, en el ejemplo que acabamos de presentar, *Juan* podría definirse como un individuo o instancia de la clase *personas* que por el simple hecho de pertenecer a esa clase presenta las mismas propiedades que el resto de miembros de la clase (como *tener primer apellido*, *tener una edad*, *tener un peso*, etc.). Evidentemente el valor de estas propiedades varía de unos individuos a otros.

A su vez, para las clases es posible definir subclases que no son más que entidades más específicas con las que comparten algunas características comunes. De esta manera, si

queremos describir la estructura de un fondo de archivo lo más sencillo sería definir las clases *Fondo*, *Sección* y *Serie*, y sus respectivas subclases *Sub-fondo, Sub-sección* y *Sub-serie*.

4.1.2. Propiedades

Otro elemento esencial dentro de la estructura de una ontología OWL son las propiedades, que permiten relacionar individuos entre sí y asignarles valores. Se distinguen básicamente dos tipos de propiedades:

- propiedades de objeto, que son las que permiten relacionar unos individuos con otros,
- propiedades de datos que relacionan individuos con un literal determinado.

Aparte de estas propiedades genéricas, OWL define una serie de características o subtipos de propiedades que enriquecen la expresividad de las mismas permitiendo especificar relaciones más complejas entre los individuos descritos en una ontología. En concreto se distinguen cuatro subtipos de propiedades:

- Las propiedades transitivas, por lo general, permiten especificar relaciones de gradación entre elementos, como por ejemplo la propiedad "*ser mayor que*".

- Propiedades simétricas son aquellas en las que si el par (*x, y*) es una instancia de la propiedad P, entonces el par (*y, x*) también lo es. Es el caso, por ejemplo, de la propiedad "*ser pariente de*".

- Las propiedades funcionales permiten definir como máximo un valor para cada individuo, como por ejemplo la propiedad "*edad*". No se imponen restricciones sobre la unicidad del valor, por lo que más de un individuo puede tomar el mismo valor para esa propiedad.

- Propiedades inversas funcionales son las que definen para una instancia un valor unívoco que no puede ser asignado a dos instancias diferentes. Es el caso de las propiedades que asignan identificadores o números personales como el ISBN a un libro o el DNI a un sujeto.

Como norma general, estos subtipos se pueden definir para cualquier tipo de propiedad aunque en el caso de las propiedades de datos solo las dos últimas tienen una aplicación clara.

4.1.3. Herencia

Otra característica importante a tener en cuenta sobre las propiedades es la capacidad que tienen las clases para dejar en herencia a sus sub-clases las propiedades que las caracterizan. Es decir, que si para la clase "*personas*", que tiene definida una propiedad "*nombre*", especificamos las subclases "*jugadores de fútbol*" y "*políticos*", estas también van a tener una propiedad "*nombre*" que identifica a sus miembros. Esta característica es aplicable a todas las propiedades independientemente de su tipología.

4.1.4. Propiedades de dominio y rango

Las clases juegan un papel determinante en el modelado de datos ya que nos permiten imponer restricciones sobre lo que se puede expresar en un aserto RDF mediante las propiedades dominio y rango. La definición de estas propiedades permite, por ejemplo, evitar la existencia de asertos del tipo:

- Natalia *es una provincia de* España
- Natalia *nació el año* azul

Tal y como vemos, ni todos los recursos ni todos los valores son válidos para definir un aserto RDF para una propiedad específica. En el primero de los casos comprobamos que el sujeto del aserto parece encajar más con una persona que con una provincia, lo cuál no resulta coherente a la vista del resto de elementos del aserto (vemos que el valor de la propiedad "*ser una provincia de*" es, en este caso, el recurso *España*). Por lo tanto, sería necesario restringir el dominio de la propiedad imponiendo la condición de que el sujeto del aserto sea un individuo que pertenezca a la clase "*provincias*" (por ejemplo, *Granada*).

En el segundo caso el sujeto del aserto es correcto (*"año de nacimiento"* es una propiedad que se puede atribuir a personas y se presupone que Natalia lo es). Sin embargo, el valor que toma la propiedad (el color *azul,* que puede ser un recurso o un literal dependiendo de la tipología de la propiedad) no parece coherente con la semántica del aserto. En este caso habría que restringir los valores que puede tomar la propiedad "*año de nacimiento*", es decir, su rango de valores. Así, por ejemplo, si consideramos que el valor adecuado para esta propiedad es un valor numérico, el rango debería restringirse a un tipo de dato numérico.

En resumen, lo que nunca hay que perder de vista es que los asertos RDF siempre van a tomar la forma *recurso-propiedad-recurso* o *recurso-propiedad-valor* y que la forma de darles coherencia semántica es realizando restricciones sobre el dominio y rango de las propiedades.

5. Creación de ontologías

Para desarrollar una ontología existen editores específicos que por lo general presentan una interfaz gráfica que permite generar la ontología de una manera sencilla y sin tener que escribir una línea de código. Existen multitud de este tipo de programas pero quizás el más utilizado es el editor ontológico *Protégé*, desarrollado por *Stanford Medical Informatics* (2010), y que permite generar ontologías de una manera bastante intuitiva.

Como recomendación general, antes de comenzar a utilizar un editor es conveniente dibujar la estructura jerárquica básica de la ontología definiendo las clases y relaciones que establecen entre sí mediante un grafo orientado (es decir, definiendo el grafo RDF de la ontología). De esta manera será más fácil identificar sus elementos básicos y plasmarlos en el editor. Una vez tenemos clara la estructura, se deberá proceder a definir de forma secuencial las clases, propiedades de objetos, propiedades de datos e instancias de las clases.

No obstante, el diseño de una ontología no es un proceso trivial y es necesario aplicar metodologías que permitan optimizar su estructura y evitar problemas comunes a todos los esquemas de conocimiento (bases de datos, tesauros u ontologías), como por ejemplo la polisemia. Este fenómeno se produce en gran medida por las diferentes conceptualizaciones

que se pueden hacer de una misma realidad en un dominio determinado y que, en un afán de economía lingüística, conviven en una única herramienta para la resolución de problemas de diferente naturaleza. Esto provoca que la eficacia de estas herramientas a la hora de acceder y recuperar información se vea seriamente mermada, por lo que se hace imprescindible definir estrategias que permitan refinar su estructura para obtener mejores resultados en términos de precisión y exhaustividad.

Para poder realizar de una manera sencilla el desarrollo, análisis y revisión de la estructura de una ontología disponemos de la metodología *Ontoclean* (Guarino y Welty 2004), la cual se basa en la definición de una serie de meta-propiedades que permiten caracterizar aspectos relevantes de la semántica de las clases y relaciones que forman una ontología para, de esta forma, detectar aquellos elementos que son superfluos o redundantes. En total son cinco meta-propiedades que imponen determinadas restricciones sobre la estructura taxonómica de la ontología:

1. La meta-propiedad "*esencialidad*" indica la necesidad de que una propiedad sea cierta para todas las instancias de una entidad determinada.

2. Para las propiedades se distinguen diferentes grados de "*rigidez*" dependiendo de si las mismas dejan de ser o no esenciales en diferentes escenarios, contextos o situaciones.

3. La meta-propiedad "*identidad*" permite establecer de manera explícita la distinción entre aquellas entidades que son iguales de las que son diferentes.

4. Por su parte la "*unidad*" identifica las partes que componen una entidad individual permitiendo detectar aquellas que son redundantes.

5. Por último, la meta-propiedad "*inclusión*" establece una serie de restricciones asociadas a la subordinación jerárquica entre clases.

En definitiva, y de un modo gráfico, podríamos decir que la aplicación de esta metodología permite "*podar*" aquellos elementos que son semánticamente prescindibles o lógicamente inconsistentes para de esta forma optimizar la estructura lógica de la ontología.

6. Mantenimiento y actualización de ontologías

Como vemos, los archivos más que sistemas de información al uso deberían ser considerados como centros de gestión de conocimiento donde los registros no son mera información, sino auténticos artefactos de conocimiento (no hay que perder de vista que la información que contienen solo adquiere su sentido completo cuando es considerada conjuntamente con el contexto de creación del registro), y donde las ontologías se muestran como herramientas que pueden ayudar a estructurarlos y organizarlos eficazmente. Por esta razón, al igual que en los archivos se aplican normas para el diseño e implantación de sistemas de información, como es el caso de la norma ISO-15489 (Llamsó Sanjuan 2009), no sería descabellado aplicar metodologías de gestión del conocimiento para explotar sus recursos de una manera más eficiente.

La gestión de conocimiento (denominada *knowledge management* o *KM* en inglés) es un concepto que engloba una serie de metodologías que permiten a una institución (como una empresa o un archivo) transferir, acumular y organizar el conocimiento y experiencia de que

disponen, ya sea a través de documentos o de la propia experiencia de las personas que trabajan en la institución, de modo que puede ser aprovechado por todos los miembros de la comunidad para resolver problemas concretos o para desarrollar su trabajo diario (Nonaka 1991) (Egbu y Boterill 2002). En el caso de los sistemas de información basados en conocimiento, donde el elemento vertebrador es un esquema de conocimiento como un tesauro u una ontología, se hace imprescindible la implantación de este tipo de metodologías de gestión para mantener adecuadamente estas estructuras y proporcionar un servicio de calidad a largo plazo.

Por lo tanto, un profesional de los archivos no sólo debería conocer qué es una ontología y cómo diseñarla, sino que sería esencial que además tuviera nociones sobre cómo implementar, gestionar y mantener estas herramientas para asegurar su funcionamiento eficiente a lo largo del tiempo. Una metodología desarrollada específicamente para gestionar sistemas basados en ontologías es *On-To-Knowledge* (Sure, Stab y Studer 2004), cuyo espíritu pretende conciliar el uso de la cultura corporativa de la organización con la identificación de los ítems de conocimiento y la aplicación de técnicas de ingeniería del software en dos fases bien diferenciadas:

1. Una primera fase de modelado de meta-procesos de conocimiento, que se centra básicamente en la puesta en marcha e implantación del sistema y su posterior mantenimiento. En esta fase de distinguen diferentes tareas genéricas orientadas a la identificación del conocimiento que va a formar parte del sistema, y que servirán como base para el desarrollo de la siguiente fase:

    - Estudio de la viabilidad del sistema.
    - Análisis de los requerimientos de la ontología para definir su estructura jerárquica básica, así como para incluir o excluir relaciones y conceptos.
    - Refinamiento de la estructura (sometiendo la ontología a procesos de poda como *OntoClean*) para obtener una ontología prototipo.
    - Evaluación de la ontología a nivel tecnológico mediante el testeo de la adecuación de su sintaxis y su consistencia semántica; a nivel del usuario, evaluando su satisfacción de uso en comparación con otras aplicaciones, y a nivel ontológico, comprobando su integridad y coherencia conceptual.
    - Implementación de la ontología en el sistema y desarrollo periódico de ciclos de evolución (que implican repetir todo el proceso descrito en este apartado) con el fin de comprobar su operatividad, y detectar y corregir posibles cambios en la estructura de conocimiento de la institución mediante las pertinentes actualizaciones en la ontología.

2. La segunda de las fases, la fase de modelado de procesos de conocimiento (ver Fig. 2), se focaliza en el análisis del funcionamiento del sistema una vez ya se ha implementado y a su "*alimentación*" con nuevos ítems de conocimiento. Engloba los procedimientos de creación e importación de registros y metadatos, el proceso de captura dinámica de nuevos ítems de conocimiento, y el acceso y recuperación de estos ítems. Los procesos se suceden de manera cíclica unos a otros de manera que la ontología se convierte en un ente "*vivo*", en constante evolución y actualización.

Figura 2. *Fase de modelado de procesos de conocimiento de On-to-Knowledge*

7. Conclusiones

El paso de los sistemas de información tradicionales a la Web se está produciendo de una manera inexorable y las nuevas tecnologías de la información juegan un importante papel como catalizadores que permiten dar el paso del formato papel a un entorno de trabajo digital.

En este trabajo hemos visto que la Archivística y los archivos deben aprovechar esta oportunidad para mejorar la descripción, el acceso, la recuperación y la gestión de los fondos. Las ontologías web se perfilan como una herramienta idónea para conseguir este objetivo, y el uso de vocabularios como OWL y metodologías como *Ontoclean* y *Onto-Knowledge* permiten diseñar, generar, depurar y mantener estas estructuras de conocimiento a lo largo del tiempo.

Por lo tanto, proponemos que los archiveros vean complementada su formación con estas nuevas tecnologías y las asuman no como un esquema rígido al que deben adaptarse, sino como una herramienta que pueden moldear de acuerdo a sus necesidades. Para ello pueden basarse en muchas de las habilidades y competencias que ya han adquirido en su formación específica como profesionales de la información (como es el caso del diseño de tesauros o el manejo del metalenguaje XML). De esta manera, la figura del archivero como descriptor y gestor del fondo se verá reconocida y reforzada como elemento clave en el funcionamiento del sistema.

Referencias bibliográficas

Antinou, G. & van Harmelen, F. (2004). *A Semantic Web Primer*. Cambridge: MIT Press.

Arana, S. & Codina, L. (2004). *La estructura conceptual de los tesauros en el entorno digital: ¿Nuevas esperanzas para viejos problemas?* Obtenido el día 10, Julio, 2010, desde http://www.lluiscodina.com/ontotesauros.doc

Beckett, D. (ed.) (2004). *RDF/XML Syntax Specification (Revised)*. Obtenido el día 13, junio, 2010, desde http://www.w3.org/TR/rdf-syntax-grammar/.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. *The Scientific American, 284* (5), 34-43

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E. & Yergeau, F. (2008). *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. Obtenido el día 13, junio, 2010, desde http://www.w3.org/TR/REC-xml/

Brickley, D. & Guha, R.V. (2004). *RDF Vocabulary Description Language 1.0: RDF Schema.* Obtenido el día 10, junio, 2010, desde http://www.w3.org/TR/rdf-schema/.

Clyde, W. & Hissaple, J.K.D. (2002). A collaborative approach to ontolgy design. *Communications of the ACM, 45* (2), 42-47.

Cruz Mundet, J.R. (2005). *Manual de Archivística*. Madrid: Fundación Sánchez Ruipérez.

Cycorp (2002). *The syntax of CycL*. Obtenido el día 14, julio, 2010, desde http://www.cyc.com/cycdoc/ref/cycl-syntax.html

Dublin Core Innitiative (2010). *Home page.* Obtenido el día 14, junio, 2010, desde http://dublincore.org/

Egbu, C.O., & Botterill, K. (2002). Information technologies for knowledge management: Their usage and effectiveness. *Journal of information technologies in construction 7*, 125-136.

Gruber, T.R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies, 43* (5-6), 907-928.

Guarino, N. (1998). Formal ontology and information systems. En N. Guarino (Ed.), *Formal* Ontology in Information Systems. Proceedings of FOIS'98 (3-17). Amsterdam: IOS Press.

Guarino, N. & Welty, C. A. (2004). An overview to OntoClean. En S. Staab & R. Studer (Eds.), *Handbook on ontologies* (151-172). Berlin: Springer – Verlag.

Hendler, J. (2001). Agents and the Semantic Web. *IEEE Intelligent Systems,* (March-April), 30-37.

Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S. Grosof, B. & Dean, M. (2004). *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*. Obtenido el día 12, junio, 2010, desde http://www.w3.org/Submission/SWRL/

ICA (International Council on Archives) (2000). *ISAD(G): General International Standard Archival Description (2nd ed.)*. Obtenido el día 15, junio, 2010, desde http://www.ica.org/en/node/30000

ICA (International Council on Archives) (2004). *ISAAR (CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families (2nd ed.).* Obtenido el día 15, junio, 2010, desde http://www.ica.org/en/node/30230

Jenz, D.E. (2003). *Bussines process ontologies: Speeding up bussiness process implementation*. Obtenido el día 31, junio, 2010, desde http://www.bptrends.com/deliver_file.cfm?fileType=publication&fileName=07-03%20WP%20BP%20Ontologies%20Jenz.pdf [Consulta: 31 junio de 2010]

Library of Congress (2002). *Encoded Archival Description (EAD)*. Obtenido el día 14, julio, 2010, desde http://www.loc.gov/ead/

Llamsó Sanjuan, J. (2009). La norma UNE-ISO 15489-1 y 2. Análisis y contenido/ Aplicación de la norma. *Arch-e: Revista andaluza de archivos, 1* (mayo), 1-17. Obtenido el día 16, junio, 2010, desde http://www.juntadeandalucia.es/cultura/archivos/impe/web_es/detalleArticulo?id=15f40f81-348e-11de-8d2f-00e000a6f9bf&idContArch=c7b697cd-2023-11de-aa97-00e000a6f9bf

McGuinness, D.L. & van Harmelen, F. (2004). *OWL Web Ontology Language Overview*. Obtenido el día 16, junio, 2010, desde http://www.w3.org/TR/2004/REC-owl-features-20040210/

Neches, R. et al. (1991). Enabling technology for knowledge sharing. *AI Magazine, 12* (3), 33-56.

Nonaka, I. (1991). The knowledge creating company. *Harvard Business Review, 69* (6), 96–104.

Palacios, J.P., Cremades, J. & Costilla, C. (2005). Towards a Web Digital Archive Ontological Unification. En *Proceedings of the Third International Conference on Information Technology and Applications (1)*, (221-226). Nueva York: IEEE Computer Society.

Palvia, S. C. J. & Sharma, S.S. (2007). E-Government and E-Governance: Definitions/Domain Framework and Status around the World. En A. Agarwal V.V. Ramana (Eds.), Foundations of E-gobernance: Proceedings of the 5th International conference in e-governance, (1-12). Obtenido el día 10, junio, 2010, desde http://www.iceg.net/2007/books/book1.html

Roberge, M. (1993). *La gestió dels documents administratius*. Barcelona: Diputació de Barcelona.

Ruiz, A.A. (ed.) (1995). *Manual de Archivística*. Madrid:Síntesis.

Standford Informatics (2010). *The Protègè Ontology Editor*. Obtenido el día 13 julio de 2010, desde http://protege.stanford.edu/

Sure, Y., Stab, S. & Studer, R. (2004). On-To-Knowledge Methodology (OTKM). En S. Staab y R. Studer (Eds.), *Handbook on ontologies* (117-132). Berlin: Springer – Verlag.

Universidad de Bremen (2008). *CASL User Manual*. Obtenido el día 14, junio, 2010, desde http://www.informatik.uni-bremen.de/cofi/wiki/index.php/CASL_user_manual

World Wide Web Consortium (W3C) (2010). *Home page*. Obtenido el día 11, junio, 2010, desde http://www.w3.org