XV Jornadas Archivísticas de la RENAIES

Archivos y Web semántica: panorama y retos profesionales

Mesa 2: Líneas de Investigación Archivística. Desarrollo teórico y su aplicación

José Manuel Morales del Castillo. El Colegio de México

e-mail: jose.morales@colmex.mx Tfno: 5449 3000 ext. 2930

INTRODUCCIÓN

La aparición de la Web como plataforma para el intercambio de información ha afectado a todos los ámbitos de la sociedad. La inexorable implantación de las tecnologías de la información está provocando que los sistemas tradicionales de información se estén moviendo hacia nuevos entornos y plataformas de trabajo, y los archivos no son una excepción.

En la década de los 90 del pasado siglo los archiveros/archivistas entendieron lo importante de dar pasos firmes hacia la presencia de los archivos en la Web y se creó el vocabulario EAD (Encoded Archival Description) (Pitti, 1999) para la descripción de guías y catálogos accesibles en línea tomando como base las normas ISAD (G) (International Council of Archives, 2011). El estándar consiste básicamente en un vocabulario definido con la sintaxis del metalenguaje XML (Extensible Markup Language) (W3C, 2003) y validado mediante una DTD (Document Type Definition) (W3C, 2002). A día de hoy, el estándar está siendo revisado y se trabaja en la versión EAD3 (Society of American Archivists, 2015) cuya principal novedad consiste en utilizar un esquema XML (XSD) (W3C, 2012b) en lugar de una DTD, ya que estas herramientas de validación han caído en desuso.

No obstante, este esfuerzo por adaptarse a los nuevos tiempos es posible que no se haya hecho a tiempo ni sea suficiente. La aparición a principios del siglo XXI de las tecnologías de Web semántica, que abren la puerta a la mejora de la descripción, acceso y

recuperación de información, obligan a revisar no sólo los estándares de descripción de recursos en la Web, sino también a adaptar sus principios teóricos y heurísticos a este nuevo entorno donde adquiere relevancia la semántica (entendida como la interrelación que se define entre agentes generadores de documentación y los propios documentos), y el contexto de creación frente a la mera enumeración descriptiva de atributos. El reto es doble desde el momento en que también se hace necesario formar a las nuevas generaciones de archivistas en el uso y aplicación de estas nuevas tecnologías, sobre todo si tenemos en cuenta que la formación que actualmente se imparte en universidades y colegios carece del componente tecnológico necesario.

En este trabajo pretendemos dar una visión de conjunto de los retos y oportunidades que estas tecnologías presentan a los profesionales de los archivos (presentes y futuros).

El resto del trabajo se estructura como sigue. En la sección 2 se hace una breve descripción de qué es la Web semántica y en la 3 repasamos brevemente el panorama que las tecnologías semánticas dibujan para la Archivística, en qué punto nos encontramos y qué tendencias se atisban en el horizonte. En la sección 4 también se apunta la necesidad de incorporar estas tecnologías al currículo de los archivistas como una reivindicar el papel del archivero en el nuevo escenario que se perfila. Por último, en la sección 5 se apuntan algunas conclusiones.

¿PERO QUÉ ES LA WEB SEMÁNTICA?

El modelo de Web semántica surge en 2001 (Berners-Lee, Hendler, & Lassila, 2001) como una extensión de la Web actual en la cual laormación está dotada de un significado definido que permite una mejor cooperación entre personas y máquinas. Dicho de otra manera, lo que se propone es conseguir que la Web funcione como una inteligencia artificial ubicua en la que una serie ante una consulta de información obtendríamos sólo aquella que necesitáramos para satisfacer una necesidad determinada (evitando de esta manera el tener que ojear miles de resultados que han sido obtenidos a partir de una correspondencia con las palabras clave usadas en nuestra búsqueda). En realidad estaríamos hablando de la misma Web que tenemos ahora pero donde el foco pasa de estar en los documentos y las palabras clave, a los datos y los conceptos interrelacionados entre sí.

El modelo se basa en dos ideas principales: el marcado semántico de recursos (lo que implica una separación formal entre el contenido y la estructura de los documentos) y el desarrollo de agentes software inteligentes capaces de procesar y operar con estos recursos a nivel semántico (Berners-Lee et al., 2001).

De una forma gráfica la WS se puede representar mediante un modelo multi-capa (fig. 1) en el que cada nivel es interoperable con los niveles vecinos (es decir, es posible intercambiar y reutilizar datos y procesos entre ellos). A grandes rasgos, podríamos decir que las diferentes capas del modelo se pueden agrupar en tres subconjuntos: las capas

inferiores, que establecen la base sintáctica del modelo; las intermedias que definen el modelo de datos y la semántica de los metadatos, y las superiores, que proporcionan los medios para establecer protocolos de seguridad y confianza.

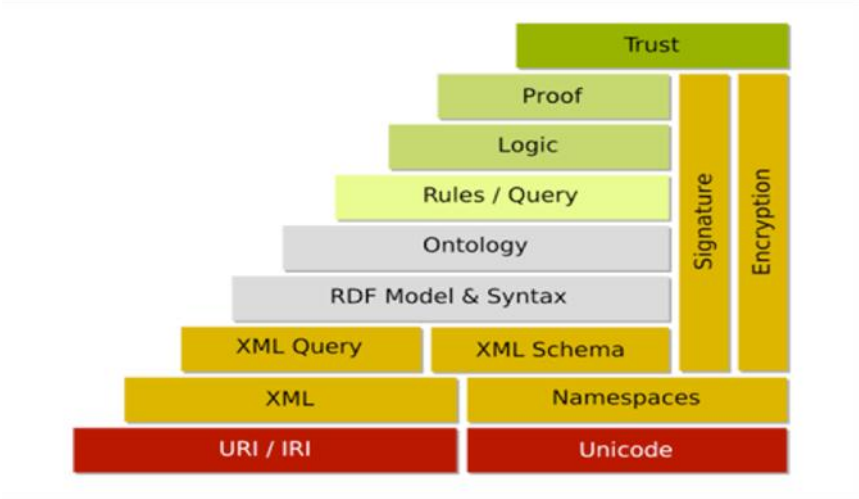

Fig. 1. Modelo multicapa de la Web semántica

El elemento vertebrador del modelo es RDF (W3C, 2014b), un lenguaje que proporciona el modelo de datos, o lo que es lo mismo, la infraestructura necesaria para codificar, intercambiar, enlazar, combinar y reutilizar metadatos estructurados. La información se organiza en asertos que toman la forma de tripletas recurso-propiedad-valor (Fig. 2) que permiten a los agentes software inferir nuevo conocimiento de los recursos de la Web utilizando para ello ontologías u otras estructuras semánticas como los esquemas conceptuales (como por ejemplo, los tesauros o las taxonomías).

Los lenguajes en los que se basa la capa semántica son RDF Schema (W3C, 2014c) y OWL (W3C, 2012a), que añaden al modelo de datos RDF elementos que permiten

explicitar con diferentes niveles de profundidad las relaciones y propiedades que se pueden definir para diferentes conceptos o entidades.

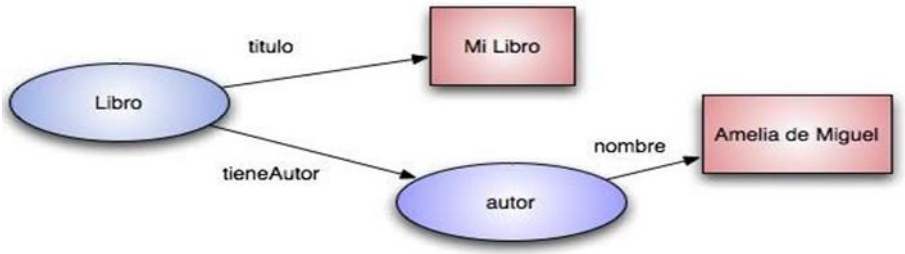

Fig. 2. Ejemplo de tripletas representadas como grafo orientado

En concreto, las ontologías en el contexto de la Web Semántica, se pueden definir como sistemas de organización del conocimiento que aglutinan una serie de conceptos relevantes del conocimiento compartido por los miembros de un dominio concreto, las relaciones que establecen entre sí estos conceptos, y los axiomas que se definen sobre estos conceptos y relaciones (Guarino, 1998). Por sí mismas las ontologías suponen un canal de comunicación entre personas y computadoras (Holsapple & Joshi, 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), de ahí que su papel dentro del esquema de la Web semántica sea clave.

LAS TECNOLOGÍAS SEMÁNTICAS Y LA ARCHIVÍSTICA

Como hemos visto, las ontologías provienen del ámbito de la inteligencia artificial y permiten representar e interrelacionar, por ejemplo, conceptos, procesos, objetos o personas para poder realizar diferentes operaciones de razonamiento con ellos. Por lo tanto, la representación de una entidad o concepto nunca se hace de manera aislada, sino que se toma en consideración la forma en que éste se relaciona con el resto entidades o conceptos que forman un dominio de conocimiento determinado. O dicho de otro modo, toda entidad concepto está caracterizado por el entorno o el contexto en el que se define y las relaciones que establece con otras entidades.

No obstante, el uso de herramientas que capturan el contexto no es algo exclusivo del área de la inteligencia artificial. En los archivos, sin ir más lejos, los cuadros de clasificación (que son elementos fundamentales en la archivística moderna) permiten estructurar lógicamente un fondo a partir del análisis funcional de las actividades que desarrolla una institución para de esta manera describir, localizar y acceder a la documentación de una manera más eficaz (que en esencia es el mismo papel que cumplen las ontologías). Entonces, y a la vista de las capacidades de las tecnologías semánticas ¿por qué no aplicarlas al desarrollo de herramientas archivísticas, donde la preservación del contexto de producción de la documentación generada por una institución es un elemento clave para la descripción, organización y gestión de los fondos?

Se podría pensar que EAD/EAD3 ya nos proporciona soluciones para el trabajo en la Web, pero hay que tener en cuenta que este vocabulario no está respaldado por una

ontología y por lo tanto no tiene la capacidad de expresar semántica, o lo que es lo mismo, no es capaz de describir el contexto de producción de la documentación y relacionarlo tanto con los agentes productores como con los atributos que caracterizan a la documentación.

Para conseguir este objetivo e implantar las tecnologías semánticas en el ámbito de los archivos se pueden explorar básicamente tres vías de desarrollo:

- Desarrollo de vocabularios semánticos respaldados por su propia ontología.
- Mapeo de vocabularios a ontologías creadas en otros ámbitos.
- Aprovechar tecnologías que ya están funcionando, como HTML5 (W3C, 2014a), para utilizarlas como contenedores de metadatos semánticos.

Ejemplo de la primera vía de implantación es el vocabulario EAC-CPF (Mazzini & Ricci, 2011) que está definido como una ontología para codificar información contextual sobre personas, entidades y familias relacionadas con materiales archivísticos, pero hasta la fecha no ha encontrado eco suficiente en la comunidad científica como para dejar de ser una solución de carácter local. En el ámbito de las bibliotecas sí que se están registrando movimientos decididos que apuesta por establecer las bases de la catalogación del futuro con vocabularios semánticos como Bibframe (Library of Congress, 2015). Este vocabulario propone una ruptura radical con el formato de catalogación MARC 21 (Library of Congress, 2007) al definir entidades y atributos comprensibles por las personas y semánticamente accesibles para las máquinas.

Sí es más sencillo encontrar en la literatura propuestas de mapeo de EAD a ontologías pertenecientes a otros dominios (asimilables de alguna manera al de la Archivística). Así, por ejemplo, encontramos el mapeo de elementos EAD a la ontología CIDOC CRM especializada en patrimonio cultural (Bountouri & Gergatsoulis, 2011), o la conversión de EAD al formato de datos enlazados de Europeana (Hennicke, Olensky, de Boer, Isaac, & Wielemaker, 2011). No obstante, no dejan de ser soluciones intermedias que no resuelven de manera integral el problema de definir un estándar para describir y relacionar semánticamente recursos archivísticos.

La otra opción posible consiste en desarrollar vocabularios específicos basados en RDFa (Resource Description Framework in attributes) (W3C, 2015b), una extensión de HTML5 que permite definir en las páginas web datos en formato RDF utilizando elementos de cualquier vocabulario semántico (como Dublin Core, por ejemplo). La idea que subyace a esta extensión es la de añadir propiedades que definen conceptos o entidades determinadas dentro de las etiquetas que permiten visualizar un documento HTML determinado. Los grandes motores de búsqueda como Google ya están aplicando estas tecnologías. En concreto utilizan Schema.org (W3C, 2015a), un vocabulario basado en RDFa que define un conjunto de tipos de recursos y sus atributos correspondientes. Desde el ámbito académico se ha trabajado en esta vía de desarrollo, pero aún no ha dado frutos tangibles.

Este panorama que en un principio no parece demasiado halagüeño, sin embargo, supone una gran oportunidad para explorar un campo en el que los profesionales de la archivística tienen mucho que decir.

EL RETO DE LA FORMACIÓN ARCHIVÍSTICA

Tradicionalmente la enseñanza de la Archivística se ha visto lastrada por un cierto inmovilismo plasmado en los planes de estudio de los centros universitarios donde se imparten estudios de Bibliotecología o Biblioteconomía y Documentación. Esto ha hecho que la disciplina se vea 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.

Este problema se traslada o quizás tiene su origen en la propia enseñanza reglada de la disciplina. Por ejemplo, si se revisa el currículum de la asignatura en algunas universidades españolas[1][2][3], tanto a nivel licenciatura como maestría, cuando se habla de enseñanza de nuevas tecnologías en Archivística la formación se reduce a explicar el uso de bases de datos en los archivos, los sistemas normalizados de intercambio de documentos,

[1] http://www.ugr.es/~epeis/docencia/archivistica/archprograma.htm
[2] http://www.cursoarchivistica.com
[3] http://www.uab.cat/web/informacion---academica---de---los---masteres---oficiales/la---oferta---de--masteres---oficiales/plan---de---estudios/plan---de---estudios/archivistica---y---gestion---de---documentos---1096480309783.html?param1=1267601207452

los sistemas de gestión de documentos y archivos electrónicos y los lenguajes de marcado como XML y EAD, pero la atención por los vocabularios semánticos es mínima o nula.

La escasa presencia en el ámbito académico de las tecnologías semánticas aplicadas a los archivos se refleja también en la poca cantidad de proyectos de investigación que se pueden encontrar en la literatura especializada, pero aún así es posible hallar algunos ejemplos. Éste es el caso del trabajo de (Palacios, Cremades, & Costilla, 2005) que presenta un sistema de gestión de archivos parlamentarios respaldados por ontologías que permiten controlar tanto la descripción de documentos (utilizando metadatos Dublin Core, y las normas de descripción archivística ISAD e ISAAR), como del dominio de trabajo del archivo para conseguir una gestión más eficaz de la documentación. Otra iniciativa es el proyecto LIAM (Linked Archival Metadata) (Morgan, 2014) cuyo objetivo es transformar los datos de la descripción archivística en datos enlazados, o el archivo sonoro de la BBC que también se basa en la aplicación de tecnologías semánticas y datos enlazados para la descripción y recuperación de ítems (Raimond & Ferne, 2013).

Y aunque los eventos científicos que tratan sobre el uso y aplicación de las tecnologías semánticas en el ámbito de los archivos aún son escasos, en los últimos años se está detectando un creciente interés en este campo. Un ejemplo es el caso del International Workshop on Semantic Digital Archives (SDAW, 2014), que comenzó a celebrarse en 2011 dentro del ámbito de la Digital Libraries Conference (JCDL, 2015).

De ahí que la formación de los archiveros en el uso de estas nuevas tecnologías no sólo supone un elemento estratégico para la propia profesión, sino que además serviría para aprovechar muchas de las habilidades y competencias que estos profesionales ya han

adquirido en su formación específica como archivistas (como es el caso del diseño de cuadros de clasificación o el manejo de 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.

CONCLUCIONES

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 y los vocabularios semánticos se perfilan como una herramienta idónea para conseguir este objetivo.

Para ello es necesario que la formación reglada de los archiveros incluya la enseñanza de este tipo de tecnologías ya que de lo contrario se corre el riesgo de que la Archivística se quede descolgada del futuro tecnológico que se nos viene encima.

BIBLIOGRAFÍA

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. Scientific American.

Bountouri, L., & Gergatsoulis, M. (2011). The Semantic Mapping of Archival Metadata to the CIDOC CRM Ontology. Journal of Archival Organization.

Guarino, N. (1998). Formal Ontology and Information Systems. In FOIS'98 (Vol. 46, pp. 3–15). doi:10.1.1.29.1776

Hennicke, S., Olensky, M., de Boer, V., Isaac, A., & Wielemaker, J. (2011). Conversion of EAD into EDM Linked Data. Retrieved May 17, 2015, from http://www.few.vu.nl/~AI.Isaac/papers/EADtoEDM.pdf

Holsapple, C. W., & Joshi, K. D. (2002). A collaborative approach to ontology design. Communications of the ACM.

International Council of Archives. (2011). ISAD(G): General International Standard Archival Description. Retrieved from http://www.ica.org/10207/standards/isadg- general-international-standard-archival-description-second-edition.html

JCDL. (2015). ACM/IEEE-CS Joint Conference on Digital Libraries2015. Retrieved May 18, 2015, from http://www.jcdl2015.org/

Jenz, D. E. (2003). Business process ontologies: Speeding up business process implementation. Retrieved May 18, 2015, from http://www.bptrends.bptrends.com/publicationfiles/07-03 WP BP Ontologies Jenz.pdf

Library of Congress. (2007). Formato MARC 21 Conciso para Datos Bibliográficos.

Retrieved May 18, 2015, from http://www.loc.gov/marc/bibliographic/ecbdspa.html

Library of Congress. (2015). Overview (Bibliographic Framework Initiative Technical Site - BIBFRAME.ORG). Retrieved May 18, 2015, from http://bibframe.org/

Mazzini, S., & Ricci, F. (2011). EAC-CPF Ontology and Linked Archival Data. In Proceedings of the 1st International Workshop on Semantic Digital Archives (SDA 2011). Retrieved from http://ceur-ws.org/Vol-801/paper6.pdf

Morgan, E. L. (2014). Linked Archival Metadata: A Guidebook. Retrieved May 18, 2015, from http://infomotions.com/sandbox/liam/tmp/guidebook.pdf

Palacios, J. P., Cremades, J., & Costilla, C. (2005). Towards a web digital archive ontological unification. In Proceedings - 3rd International Conference on Information Technology and Applications, ICITA 2005 (Vol. I, pp. 221–226).

Pitti, D. V. (1999). Encoded archival description: An introduction and overview. D-Lib Magazine. doi:10.1045/november99-pitti

Raimond, Y., & Ferne, T. (2013). World Service Radio Archive Prototype. Retrieved May 18, 2015, from http://worldservice.prototyping.bbc.co.uk/

SDAW. (2014). Semantic Digital Archives Workshop 2014. Retrieved May 18, 2015, from http://sda2014.dke-research.de/

Society of American Archivists. (2015). EAD3 Status Update | Society of American Archivists. Retrieved May 17, 2015, from http://www2.archivists.org/groups/technical-subcommittee-on-encoded-archival- description-ead/ead3-status-update

W3C. (2002). XHTML 1.0: The Extensible HyperText Markup Language (Second Edition). Retrieved May 17, 2015, from http://www.w3.org/TR/xhtml1/#dtds

W3C. (2003). Extensible Markup Language (XML). Retrieved May 17, 2015, from http://www.w3.org/XML/

W3C. (2012a). OWL 2 Web Ontology Language Document Overview (Second Edition). Retrieved May 18, 2015, from http://www.w3.org/TR/owl2-overview/

W3C. (2012b). W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures. Retrieved May 17, 2015, from http://www.w3.org/TR/xmlschema11-1/

W3C. (2014a). HTML5. Retrieved May 18, 2015, from http://www.w3.org/TR/html5/

W3C. (2014b). RDF - Semantic Web Standards. Retrieved May 18, 2015, from http://www.w3.org/RDF/

W3C. (2014c). RDF Schema 1.1. Retrieved May 18, 2015, from http://www.w3.org/TR/rdf-schema/

W3C. (2015a). Home - schema.org. Retrieved May 18, 2015, from https://schema.org/

W3C. (2015b). RDFa Core 1.1 - Third Edition. Retrieved May 18, 2015, from

http://www.w3.org/TR/rdfa-syntax/