Autor: Daniel Martínez Ávila – dmartine@bib.uc3m.es
Profesor: Tomás Nogales Flores - nogales@bib.uc3m.es
Tratamiento de documentos digitales con tecnologías XML
Master en Investigación en Documentación 2007-2008

Tendencias y usos de XML en Biblioteconomía y Documentación

Señalaba Kyle Banerjee en su artículo del 2002 "How does XML help libraries?" (1) que eran muchas las innovaciones tecnológicas que habían creado fuertes expectativas en el mundo de las bibliotecas, pero pocas lo habían hecho con la fuerza de XML. De esta forma se ha dicho que aplicaciones en Java para el mundo bibliotecario o el mismísimo protocolo Z39.50 iban a revolucionarlo todo, incluso mucho antes de que hubiera nada implementado en estas tecnologías o herramientas, y que sin embargo poco tendrían que ver estos casos con lo que se ofrecería en el supuesto de una plena implantación del XML. Lo cierto es que, con la experiencia ya contrastada por el uso y la aceptación, se ha demostrado que dicho lenguaje (o mejor dicho gramática) no ha sido capaz de eliminar como se prometía todos los problemas existentes de formatos y conversiones de tipos de datos, ni tampoco ha sido un lenguaje capaz de integrar todos los recursos de información existentes, ni por supuesto se ha convertido en la panacea que librara el mundo bibliotecario de todos los problemas de estandarización que nos atañen. Pero en este caso el problema de la relativa falta de éxito no ha sido la falta de potencial, sino como muy bien señala Banerjee en su artículo, ha sido por falta de tiempo o dinero a la hora de liberar todas las posibilidades en forma de aplicaciones concretas, un defecto que por otra parte como muy bien se puede intuir, se corregiría principalmente a través del uso e impulso de esta nueva herramienta por parte de todos los profesionales de la Documentación.

El problema de la estandarización en las bibliotecas ha crecido exponencialmente en los últimos años. Con la instauración definitiva de los recursos de Internet entre los grandes activos de los centros, el concepto de catálogo como herramienta de acceso a un conjunto integrado de recursos físicos se ha ido difuminando. Hoy en día entre los servicios que se ofertan en cualquier centro también nos encontramos revistas electrónicas, páginas web, bases de datos, libros electrónicos y una serie de recursos para los que no estaba preparado el catálogo, como consecuencia de esto muchas veces se ha optado por no incluirlos en él, y por ende, entre los usuarios ha decrecido el acceso, o incluso conocimiento, de dichos recursos. Uno de los aportes principales del XML es la posibilidad de aunar, por medio de la estandarización, el acceso e intercambio a muchos tipos de registros, una cuestión que no era posible, por ejemplo, con el uso de MARC de forma aislada y que con la ayuda de XML sí que podría alcanzarse. Otra de las ventajas de XML es la posibilidad de presentar la misma información para diferentes perfiles de usuarios.

La tecnología XML puede ser muy flexible y versátil, muchas veces consiguiendo esto a cambio de un coste de eficiencia computacional. Es por esta razón por la que XML no es la opción más adecuada para aquellos casos que se podrían solucionar a través bases de datos simples y estructuradas, una solución para la que quizás no haría falta tanto potencial. Así nos encontramos ejemplos que se pueden hacer y otros ejemplos que se podrían hacer pero no hace falta usar (aunque se podría) una herramienta como XML. Entre los primeros, y como representación de las últimas

tendencias en materia de aplicación a la Biblioteconomía y Documentación nos encontramos los siguientes ejemplos:

- Tecnología XML en los Sistemas de Gestión de Contenidos (2). En un Sistema de Gestión de Contenidos la tecnología XML permitiría la identificación de los contenidos dentro de los documentos digitales, lo cual permitiría su reutilización en cualquier momento y en cualquier circunstancia. Tal como se señala en el *white paper* de la compañía Documentum sobre XML y la gestión de contenidos "que las organizaciones puedan intercambiar datos y contenidos empresariales de forma que cada una de ellas pueda entender esta información y utilizarla en sus procesos de negocio". En cuanto a la publicación multicanal y multiformato se podría abordar este problema mediante la sindicación de contendidos, como por ejemplo el uso de RSS. Y por último para aquellos objetos digitales que no pueden ser tratados con XML no hay que olvidar que sí se podrían tratar descripciones de ellos, así como el establecimiento de relaciones que permitan una identificación y posterior recuperación de los mismos. El uso de metadatos también cobraría un papel importantísimo para este modelo de gestión de contenidos.

- Aplicación de XML orientado al texto (3). Como se ha comentado anteriormente, las posibilidades de XML no se aprovechan al máximo cuando se trabaja con datos muy estructurados como por ejemplo los provenientes de formularios o similares. Lo cierto es que el verdadero potencial de dicho lenguaje de marcado lo encontramos al trabajar sobre texto, situación que ve fundamentada originalmente desde los objetivos de su más inmediato antecesor, SGML. De esta forma algunas de las iniciativas para tratamiento de texto como TEI o EAD tienen un origen común en el entorno de SGML y muestran una evolución paulatina hacia su incorporación en sistemas XML. TEI (Text Encoding Initiative, o Text Encoding and Interchange, según fuentes) es un estándar internacional que según sus desarrolladores "ayuda a bibliotecas, museos, editores, eruditos e investigadores a representar todo tipo de textos literarios y lingüísticos para la investigación y la enseñanza en línea, usando un esquema de codificación que es máximamente expresivo y mínimamente obsolescente", en definitiva, se busca un estándar independiente de hardware y software que permita codificar e intercambiar textos en cualquier lenguaje natural, de cualquier fecha, tipo o género literario sin restricción sobre su forma o contenido. De dicha iniciativa surgieron una DTD y unas directrices de uso libre desarrolladas para dar soporte XML. En cuanto a EAD (Encoded Archival Description) lo que se intentó es construir "un estándar para codificar instrumentos de descripción archívistica por medio de SGML" y XML. Entre los requisitos de sus orígenes debían estar "la capacidad de presentar información descriptiva extensiva e interrelacionada, preservar las relaciones jerárquicas existentes entre niveles de descripción, representar la información descriptiva heredada por un nivel jerárquico desde otro, o permitir la indización y recuperación de un elemento específico. A partir de ahí se crearon DTDs con apéndices tan útiles como ejemplos de uso o pasarelas a otros estándares como MARC21. También es objeto de mención el hecho de que el desarrollo de EAD se vea acompañado por la participación de su comunidad de usuarios, que propone y mejora aspectos como por ejemplo la compatibilidad con la norma ISAD (G) u otros. En definitiva, tanto TEI como EAD, como cualquier otra

aplicación de lenguaje de marcado a texto, lo que proporciona es un nuevo horizonte y universo de posibilidades a la hora de intercambiar y acceder a los documentos en un entorno electrónico.

- Estándares de documentación en XML para el desarrollo del gobierno digital (4). Se han llevado diversas iniciativas en el desarrollo de estándares para el intercambio de documentos digitales e interoperabilidad en el contexto del llamado gobierno digital. En estos ambientes de gobierno digital, así como en la llamada administración electrónica, no sería posible una utilización tan eficiente y libre de la documentación si no se buscara la adopción de tecnologías XML, por todas las razones expuestas anteriormente de normalización, reutilización, procesamiento y más. Ejemplos concretos de este caso ya implantado los encontraríamos en países como Australia, Canadá, Chile, Dinamarca, Estados Unidos, Finlandia, Hong Kong, Nueva Zelanda, Reino Unido o Singapur.

Además de todos estos trabajos también se están publicando numerosas tesis doctorales en Biblioteconomía y Documentación relacionadas con las posibilidades de uso del XML, lo que no deja de ser otro ejemplo más de la buena salud que goza dicha tecnología dentro de nuestro ámbito. Todas las tesis reseñadas a continuación fueron presentadas en la Universidad Carlos III de Madrid en un menos de una década:

- Tratamiento y difusión en Internet de información jurisprudencial mediante tecnologías XML: aplicación al caso del tribunal constitucional (5). Estudio llevado por Bonifacio Martín Galán en el 2001 en el que se demuestra "la utilidad de las tecnologías XML para la construcción de bases de datos juridisprudenciales que han de ser publicadas y consultadas a través de la red Internet". En esta ocasión la utilidad expuesta consistiría en "garantizar un correcto tratamiento y análisis de la información contenida en los documentos jurídicos, describiendo tanto su forma como su contenido, desarrollando lenguajes controlados capaces de reducir la ambigüedad de los mismos y generando una serie de subproductos documentales para la localización y recuperación de la información jurídica de interés". En este caso estaríamos ante una aplicación del lenguaje de marcado de texto en la que tratando semánticamente, y formateando las Sentencias del Tribunal Constitucional contenidas en una base de datos, se permite una presentación y recuperación óptima para unos usuarios que accederían desde Internet.

- La organización hipertextual del ordenamiento jurídico por medio de tecnologías XML: aplicación a la normativa del IRPF (6). Estudio llevado por Carmen Arellano Pardo en el 2003 en el que también se mostraban las ventajas aportadas a la recuperación de legislación por medio de organización hipertextual. En dicho estudio se consideraría como unidad de información al conjunto de normas, desglosándose en otras unidades inferiores como la norma o el artículo, las cualidades semánticas entre unas unidades y otras, así como las de la propia estructura lógica del documento, serían descritas gracias a las características del XML. Dicha tecnología también permitiría por ejemplo el acceso tanto desde la norma o elemento que referencia como desde la que es referenciada, es decir, una navegación en ambos sentidos optimizando de forma invisible la aplicación para el usuario. Otra de las ventajas del uso del XML al ordenamiento jurídico es la posibilidad de reutilización de información, con las consiguientes ventajas

de eficiencia y coherencia para su presentación. También en este estudio se propone la creación de una DTD que fuera aplicable a los documentos según las Directrices sobre la forma y estructura de los anteproyectos del Ley, tanto para los documentos ya existentes como una nueva DTD que obligara a los documentos que surgieran a seguir dichas directrices.

- Aplicaciones de XML para la documentación periodística: efectos sobre los centros de documentación de prensa (7). Estudio llevado a cabo por David Rodríguez Mateos en el 2003 en el que se demostraba la utilidad de la aplicación de tecnologías XML a la documentación periodística. En los nuevos tiempos informativos de prensa digital y publicaciones en línea se propone el uso de una tecnología que se adapte a la nueva realidad, un tratamiento en el que "se requiere menos énfasis en el análisis formal y en el almacenamiento, y más en el análisis de contenido y en la difusión/recuperación del mismo". Este tipo de documentos muchas veces son conjuntos de referencias nominales a objetos, es decir, con un conjunto de objetos se puede crear múltiples documentos sin necesidad de modificar ningún objeto, es aquí donde entra en juego la ventaja de reutilización que proporciona la tecnología XML. En el estudio también se exponen algunos ejemplos concretos de uso de lenguaje de marcado como PRISM, RSS, NewsML, NITF, SportsML o PAM y se advierte que todos los lenguajes son combinables entre sí.

- Una aproximación a la reutilización de estructuras documentales definidas con XML en el marco de la web semántica (8). Estudio elaborado por Elena Castro Galán en el 2003 en el que se enfatiza las posibilidades de reutilización que ofrece XML aplicadas a la web semántica. En dicho estudio primero se analizan las propuestas más interesantes de modelado de información en web como: DTDs, XML, XML Schemas, RDF y RDF Schemas, todo ello acorde con las recomendaciones del W3C. Básicamente lo que se pretende es una búsqueda de métodos de almacenamiento y recuperación eficientes aplicando dichas tecnologías a determinadas colecciones de documentos. En el estudio también se trabaja en la utilización de un metaesquema basado en un modelo conceptual (como podría ser el Modelo Relacional), para realizar la conversión entre esquemas web y esquemas lógicos, así como la construcción de un repositorio para gestionar este tipo de esquemas, más concretamente basados en DTDs, aunque podría extenderse a otro tipo de esquemas.

Además de todos los ejemplos de aplicación expuestos anteriormente, es fácil encontrar en la literatura especializada ejemplos o propuestas nuevas de uso de todo tipo. Así encontramos obras dedicadas a la aplicación de la tecnología XML a las bibliotecas (9) en la que se nos expone la ventaja de permitir al software de ordenador procesar gran cantidad de campos bibliográficos y objetos de forma que se puedan seleccionar partes, reutilizar o producir diferentes presentaciones. En esta obra también se dividen las posibles aplicaciones del XML en siete: catálogos bibliotecarios, préstamo interbibliotecario, catalogación e indización, desarrollo de colecciones, bases de datos, migración de datos e interoperabilidad entre sistemas. En concreto se habla de aplicar por ejemplo XML a MARC o Z39.50 para mantener el catálogo y recuperar información respectivamente. También se habla de hacer interoperables objetos digitales gracias al uso de METS (Metadata Encoding and Transfer Standard).

Otra obra (10) de similares características a la anterior, centrada sobre todo en la mejora del acceso y gestión de información, hablaría de registros bibliográficos y autoridad proponiendo y analizando la creación de schemas para MARC o AACR, culminando finalmente con XOBIS (XML Organic Bibliographic Information Schema) que fue desarrollado en septiembre del 2002 y que superaría algunos de los problemas presentados por los anteriores sistemas.

En definitiva, se puede comprobar a través de la literatura especializada una esplendida salud de la tecnología XML aplicada al mundo de la Biblioteconomía y Documentación, con numerosos proyectos, investigaciones y posibilidades abiertas que surgen y surgirán entre los profesionales e investigadores de nuestro campo. Simplemente es cuestión de tiempo ver como poco a poco ese gran ideal aparentemente inalcanzable de eliminar todos los problemas existentes de formatos y conversiones de tipos de datos pueda ser alcanzado entre todos. Hasta entonces nosotros, como profesionales e investigadores especializados, sólo podemos hacer una cosa para conseguirlo: desarrollar, investigar y hacer visibles en la medida de lo posible todas las virtudes de ese infravalorado y versátil metalenguaje llamado XML.

Referencias:

1. BANERJEE, Kyle. *How does XML Help Libraries?* , 2002.

2. MARTÍN GALÁN, Bonifacio; and ARELLANO PARDO, María del Carmen. La información en la Posmodernidad: la Sociedad del Conocimiento en España e Iberoamérica Madrid: Centro de estudios Ramón Areces, 2004. *Los Sistemas De Gestión De Contenidos: Un Paso Más Hacia La Gestión Integral De La Información y El Conocimiento En Las Organizaciones*. ISBN 8480046430.

3. NOGALES FLORES, Tomás; and MARTÍN GALÁN, Bonifacio. La información en la Posmodernidad: la Sociedad del Conocimiento en España e Iberoamérica Madrid: Centro de estudios Ramón Areces, 2004. *Dos Campos De Aplicación De XML Orientado Al Texto: La Literatura (TEI) y La Descripción Archivística (EAD)*. ISBN 8480046430.

4. BAEZA-YATES, Ricardo; and VÁSQUEZ, Cristian. La información en la Posmodernidad: la Sociedad del Conocimiento en España e Iberoamérica Madrid: Centro de Estudios Ramón Areces, 2004. *Estándares De Documentación En XML Para El Desarrollo Del Gobierno Digital*. ISBN 8480046430.

5. MARTÍN GALÁN, Bonifacio; and NOGALES FLORES, Tomás [dir.]. *Tratamiento y Difusión En Internet De Información Jurisprudencial Mediante Tecnologías XML: Aplicación Al Caso Del Tribunal Constitucional : Tesis Doctoral.* Getafe: Universidad Carlos III de Madrid, Departamento de Biblioteconomía y Documentación, 2001.

6. ARELLANO PARDO, María del Carmen; and NOGALES FLORES, Tomás [dir.]. *La Organización Hipertextual Del Ordenamiento Jurídico Por Medio De Tecnologías XML, Aplicación a La Normativa Del IRPF : Tesis Para La Obtención Del Título De Doctor En Documentación.* Getafe: Universidad Carlos III de Madrid, Departamento de Biblioteconomía y Documentación, 2003.

7. RODRÍGUEZ MATEOS, David; and HERNÁNDEZ PÉREZ, Antonio [dir.]. *Aplicaciones De XML Para La Documentación Periodística : Efectos Sobre Los Centros De Documentación De Prensa : Tesis Doctoral.* Getafe Madrid: Universidad Carlos III de Madrid, Departamento de Biblioteconomía y Documentación, 2003.

8. CASTRO GALÁN, Elena; NOGALES FLORES, Tomás [dir.] and VELASCO DE DIEGO, Manuel [dir.]. *Una Aproximación a La Reutilización De Estructuras Documentales Definidas Con XML En El Marco De La Web Semántica : Tesis Doctoral.* Getafe Madrid: Universidad Carlos III de Madrid, Departamento de Biblioteconomía y Documentación, 2003.

9. *XML in Libraries Roy Tennant (Ed.).* New York: Neal-Schuman Pub, 2002. ISBN 1555704433.

10. MILLER, Dick R.; and CLARKE, Kevin S. *Putting XML to Work in the Library : Tools for Improving Access and Management.* Chicago: American Library Association, 2004. ISBN 0-8389-0863-2.