XML, ¿UNA INFRAESTRUCTURA PARA LA BIBLIOTECA DIGITAL?: EL PROYECTO COVAX

Francisca Hernández y Carlos Wert
Residencia de Estudiantes

Este artículo plantea algunas cuestiones relacionadas con el papel del lenguaje XML (Lenguaje de marcación extensible) en relación con la construcción de bibliotecas digitales. Su alcance es limitado y se refiere específicamente a la experiencia de un proyecto de ámbito europeo, desarrollado bajo el liderazgo de la Residencia de Estudiantes, cuyas conclusiones, no obstante, aspiran a cierto grado de generalidad.

Una institución de memoria

Las peculiaridades de la Residencia de Estudiantes empujaban a esta institución por el camino de la innovación en la prestación de servicios de documentación. No es una biblioteca, un archivo ni un museo, sino lo que, a impulso de la Comisión Europea, cada vez más gente denomina una institución de memoria[1], cuya colección incluye documentos de todo tipo. Reúne además, a esta condición, la de un activo centro creador y transmisor de cultura y, sin ser un típico centro de investigación, realiza una amplia actividad de I+D+I.

Archivo virtual

Cuando la Residencia se planteó crear y ofrecer en la red una colección digital, cuando inició su transformación en *biblioteca digital* (o, mejor, en biblioteca híbrida, que ofrece recursos digitales junto a los analógicos) escogió, para nombrar a la colección digital que creó junto a otras instituciones (y a la página web que permite acceder a ella), el título *Archivo virtual de la Edad de Plata*[2]. El nombre llamaba la atención sobre cierto protagonismo de los documentos de archivo, pero igualmente podría haberse hablado de biblioteca virtual (la base de datos incluye un número importante de referencias de monografías y publicaciones periódicas), poniendo de manifiesto la inexistencia de un término acuñado que abarcase ambas realidades.

Y este era el primero de sus rasgos innovadores en relación con los sistemas de información al uso: su base de datos incluía documentos de todo tipo (libros, revistas, manuscritos, correspondencia, material gráfico, objetos artísticos…). Contenía tanto entradas de catálogo bibliográfico como inventarios de colecciones documentales o artísticas. Pretendía, desde el primer momento, subrayar la unidad que, para el usuario, representa la colección de cualquier institución de memoria y la conveniencia de que los mecanismos de búsqueda y recuperación de información obvien la necesidad de acceder a bases de datos distintas que describen contenidos distintos con procedimientos distintos[3].

En segundo lugar, reunía en un mismo sistema de información a varias instituciones con colecciones relacionadas, creando una colección virtual de los documentos pertinentes para la reconstrucción de la unidad de la memoria del periodo de la cultura española que identificamos como la Edad de Plata[4].

---

1 Wert, Carlos. Las instituciones de memoria e Internet, incluido en José Antonio Millán y otros, eds., Telecomunicaciones, sociedad y cultura, Madrid, 2002, pp. 93-102.

2 http://www.archivovirtual.org.

3 Hernández, Francisca y Xavier Agenjo. ¿Tres vías al conocimiento? La información de archivos, bibliotecas y museos y el derecho de los ciudadanos a los documentos primarios en Información y derecho de los ciudadanos: La confrontación entre teoría y realidad en el 20º aniversario de la Constitución: Actas del VII Congreso Nacional de ANABAD. - En: *Boletín ANABAD*. - n. XLIX (1999) n. 3-4, Julio-Diciembre, p. 559-567.

4 Wert, Carlos. Las tecnologías de la información y el rescate del patrimonio del exilio, ponencia presentada en el Seminario La Numancia errante (Valencia, junio de 2001), en prensa.

En tercer lugar, junto a las descripciones de las piezas documentales, daba acceso, cuando los titulares de los derechos de los documentos así lo querían, a facsímiles digitales de los propios documentos (y, en ciertos casos, a transcripciones textuales de los mismos).

Tres años después de su inicio, el *Archivo virtual* da acceso a mas de 200.000 documentos pertenecientes a 60 colecciones y ha digitalizado más de 750.000 imágenes. Ha representado pues un notable esfuerzo en digitalización del patrimonio cultural y también una tentativa de ofrecer a los usuarios un acceso integrado, global a colecciones físicamente separadas, pero relacionadas por su contenido.

Algunas carencias

Pero, desde el comienzo, a la vez que sus capacidades, el *Archivo virtual* mostró ciertas insuficiencias. En primer lugar, las limitaciones que presenta el formato MARC para la descripción jerárquica de documentos a varios niveles (series, subseries, etc.), necesaria para un tratamiento y recuperación adecuados de los inventarios de archivo. Es decir, el sistema de recuperación se comportaba correctamente con registros bibliográficos o con descripciones de archivo de primer nivel, pero no respondía satisfactoriamente a la presentación contextualizada de información correspondiente a series, subseries y unidades documentales.

En segundo lugar, el acceso a las colecciones de los centros participantes se solucionaba por medio de un catálogo colectivo centralizado en la Residencia de Estudiantes y no como una verdadera red de bases de datos distribuidas. En tercer lugar, la gestión de imágenes se realizaba por un procedimiento rudimentario que apenas permitía otra cosa que hacerlas accesibles por Internet. Era necesario disponer de una herramienta más compleja para la gestión de las imágenes digitales que permitiera su paginación; asociarles información complementaria como la estructura de índices, capítulos, epígrafes, etc., para facilitar la navegación en su interior; la gestión de permisos de visualización, impresión y descarga; y el control de su uso conforme a los requerimientos de las instituciones propietarias de los documentos y de los titulares de los derechos intelectuales.

Por otra parte, la evolución de los buscadores de información había convertido en obsoleto el mecanismo de búsqueda por comparación exacta de patrones de palabras. Este mecanismo manifestaba sus limitaciones aún más con la aparición de los buscadores más modernos para ajustan la relevancia de los resultados encontrados (*Google* ha modificado las expectativas respecto al comportamiento de un buscador). Por otro lado, la recuperación de información estaba también limitada por la falta de herramientas de reordenación, presentación o reutilización de los resultados.

Apuesta por XML

La solución a algunos de estos problemas mayores se concibió como un proyecto de investigación. Este es el origen del proyecto COVAX (*Contemporary Culture Virtual Archives in XML*[5]). Preparado en 1999, en paralelo a la construcción del *Archivo virtual*, el proyecto se presentó a la primera convocatoria del programa IST (Tecnologías para la Sociedad de la Información) del Quinto Programa Marco europeo de I+D y obtuvo el apoyo de la Comisión Europea. Y, en el propio título del proyecto, aparecía, como una opción radical, el lenguaje XML.

[5] http://www.covax.org. En el proyecto han participado las siguientes organizaciones: Residencia de Estudiantes, Software AG España, Biblioteca de Menéndez Pelayo y Universitat Oberta de Catalunya (España), Angewandte Informationstechnik Forschungsgesellschaft m.b.Hm y Salzburg Research Forschungsgesellschaft m.b.H.(Austria); ENEA (Italia); Blekinge Tekniska Högskola (Suecia) y LASER y South Bank University (Reino Unido). Un balance del proyecto puede consultarse en Yeates,Robin: An XML infrastructure for archives, libraries and museums: resource discovery in the COVAX project. En: *Emerald*, vol 36, n. 2 (2002): 72-82, accesible en http://www.emeraldinsight.com/0033-0337.htm.

Hay que mencionar que, en el momento de la propuesta del proyecto, XML estaba considerado el lenguaje del futuro para el intercambio de información en Internet, pero aún no había comenzado el desarrollo de todo su potencial y tanto los códigos disponibles para tipos de documentos específicos como inventarios de archivo, textos electrónicos o registros bibliográficos, como el software (editores, parsers y bases de datos) se encontraba en un estado poco desarrollado. Para los archivos y bibliotecas, XML representaba entonces un futuro prometedor, pero sus realizaciones eran limitadas.

Alcance de COVAX

El proyecto se desarrolló a lo largo de los años 2000 y 2001, y perseguía cuatro objetivos principales, que se correspondían con los objetivos del *Archivo Virtual* y de otros socios del proyecto. En primer lugar, como se ha dicho, crear un mecanismo de acceso global al patrimonio cultural, intelectual y científico conservado en archivos, bibliotecas y museos, evitando que las diferencias de tratamiento, proceso y descripción se traslucieran en los sistemas de búsqueda y recuperación de información. En segundo lugar, la explotación a través de Internet de los recursos informativos realmente existentes en las instituciones de memoria, es decir, incrementar su accesibilidad haciendo visible y recuperable su contenido. En tercer lugar, se trataba de diseñar un sistema de recuperación que interconectara archivos, bibliotecas y museos en un entorno distribuido. Por último, se optaba por la aplicación estricta de normas en el campo de la estructura y la recuperación de información. Para todo ello, el eje común era la utilización de XML. El proyecto quería responder a la pregunta ¿es viable la utilización de XML para codificar y recuperar información de archivos, bibliotecas y museos en bases de datos distribuidas?

El formato MARC y el protocolo Z39.50 ya habían probado su eficacia en la recuperación de información bibliográfica en distintas bases de datos al mismo tiempo. COVAX se diseñó para, a través de la construcción de un prototipo, poner a prueba la utilidad de XML para realizar esta misma tarea y sacar a la luz a lo largo del desarrollo del proyecto los problemas que esto comportaba y los riesgos a los que había que hacer frente en su aplicación, tanto en la descripción de documentos como en su recuperación.

Reutilizar lo existente

La filosofía del proyecto estaba basada en que las bibliotecas, archivos y museos han ido creando a lo largo de sus años de actividad innumerables instrumentos de descripción y control (desde listados a bases de datos caseras sobre los que han basado gran parte de su actividad informativa), y que habitualmente no disponen de recursos para rehacer este trabajo. Por tanto, también XML debía servir para codificar esos instrumentos ya existentes y dar el salto que hiciera disponible esa información en Internet. Si se trataba de situar a archivos, bibliotecas y museos en Internet, en el mundo del intercambio de información, y hacerlo de una manera sostenible e independiente, sólo podía hacerse desde el supuesto de que a partir de la conversión de los datos originales a XML se dispondría de un medio de codificar información de forma normalizada. Se pretendía también no repetir la experiencia del formato MARC (una babel de *dialectos*): no contribuir a la *babelización* de XML creando nuevas definiciones de tipos de documentos demasiado adaptadas a los contenidos disponibles.

Por ello el plan de trabajo de COVAX se inició con el análisis de los sistemas de información de los participantes como mecanismo para evaluar los procedimientos necesarios para la conversión en todos sus aspectos (desde el uso de formatos locales a los juegos de caracteres) y para definir el conjunto de los documentos que debían formar parte del prototipo. En esta fase resultó crucial el conocimiento de la estructura de la información, de los puntos de acceso convenientes y el estudio sobre el estado del arte en relación con XML (otros proyectos, editores, sistemas de gestión de bases de datos, etc.[6]) No obstante, éste varió considerablemente a lo largo del proyecto (se produjeron múltiples novedades en productos, versiones… y,

en algún caso, las capacidades de las herramientas en el momento en que fueron necesarias (o su evolución) supusieron alguna limitación para el desarrollo del mismo.

Conversión

Los procesos de conversión[7] tuvieron características diferentes en lo que se refiere a los registros bibliográficos y a los inventarios de archivos o catálogos de museos. En el caso de los textos electrónicos, y en muchos inventarios, no puede hablarse de conversión puesto que la codificación en TEILite, versión abreviada de la DTD (definición de tipo de documento) TEI (Text Encoding Initiative: DTD para textos electrónicos) o en EAD (Encoded Archival Description: DTD para descripciones de archivos) se realizó directamente. En el caso de los registros bibliográficos, supuso pasar de los distintos formatos de intercambio ISO 2709 (UKMARC, IBERMARC, CATMARC y UNIMARC) a MARC 21, formato base para la correspondiente definición de tipo de documento, la MARC DTD. Para muchos de los socios del proyecto, esta conversión ha supuesto un paso importante en la normalización de sus bases de datos bibliográficas independientemente del uso de XML, puesto que una de las salidas de información ha sido la creación de registros MARC 21. También son de interés para cualquier institución que disponga de registros y descripciones en bases de datos caseras (o no tan caseras) las instrucciones para transformar registros de MSAccess a XML (y por tanto a MARC 21) que se han desarrollado en el curso del proyecto. Este proceso de conversión produjo la creación de 4 bases de datos de documentos EAD, 8 de descripciones MARC, 3 bases de datos con descripciones AMICO y 2 bases de datos con textos en TEI.

Prototipo

El prototipo[8] desarrollado por el proyecto, del que se han producido dos versiones, está formado por los servidores de bases de datos de los socios y el servidor del sistema, constituido por tres componentes que, juntos, forman el núcleo del sistema:

- Un interfaz multilingüe para el acceso a todas las funcionalidades del sistema en los cinco idiomas de los socios (español, catalán, sueco, inglés, alemán e italiano) basándose en el uso de hojas de estilo multilingües. Ofrece además una base de datos (al modo del servicio *explain* de Z39.50) con una descripción de las instituciones, los servidores y el contenido de sus bases de datos.
- Un meta-buscador que gestiona la interacción entre todas las bases de datos individuales y resuelve las consultas globales. Desde el primer momento de la definición del sistema se consideró de crucial importancia la implementación del modelo Z39.50 (protocolo de búsqueda en bases de datos distribuidas). Para ello se han utilizado las reglas de codificación XER (XML Encoding Rules), ideadas como mecanismo para la interoperabilidad entre sistemas Z39.50 y sistemas Web/Internet.
- Herramientas de administración, componentes intermedios que relacionan y transmiten la información entre la interfaz y el motor de búsqueda.

Resultados

6 El resultado de este estudio puede verse en http://www.covax.org/covax_e/public_docum/p_documents.htm

7 Hernández, Francisca; Peter Linde; Bob Mulrenin and Robin Yeates. Converting heterogenous cultural catalogues and documents to XML. Strategies and solutions of the COVAX project. *Proceedings of the International Conference on Electronic Publishing 2001* (ELPUB 2001), celebrado en Canterbury, UK, 5-7 July 2001, cuyas actas han sido publicadas por IOS Press (http://www.iospress.nl/).

8 http://www.covax.at

En primer lugar, se han demostrado ampliamente los objetivos marcados en el inicio del proyecto, confirmándose la viabilidad del uso de XML para facilitar el acceso a bases de datos distribuidas de archivos, bibliotecas y museos. Desde el principio del proyecto COVAX planteó su conformidad con la normativa de los distintos ambientes profesionales en los que se movía. Su objetivo no era el de crear normas nuevas sino contribuir a la aplicación de las ya existentes. No se trataba, por tanto de crear nuevas DTDs para cada uno de los sistemas de información de los socios, especialmente en caso de las descripciones bibliográficas, sino de normalizar estas descripciones existentes conforme a las normas más sólidas y que dispusieran de un mayor aparato complementario de guías, directrices, software gratuito, etc.

Sin embargo, se aprecia que es necesaria la evolución de algunas DTDs en varios sentidos. La MARC DTD, que es en la actualidad una transposición del formato MARC21 es susceptible de adoptar un modelo mucho más dependiente de las normas de catalogación del tipo ISBD o *Functional Requirements for Bibliographic Records*[9]. Es necesario realizar adaptaciones en la presentación de documentos, sobre todo TEI y EAD, que pueden ser excesivamente grandes para ser manejados, con unos tiempos de respuesta adecuados, por las bases de datos, motores de búsqueda y herramientas de administración. El establecimiento de puntos de acceso e índices en las bases de datos XML se ve dificultado, en los casos de TEI y EAD, por el alto nivel de anidamiento de sus elementos. Es de esperar que la evolución del software que soporta XML, sobre todo en bases de datos, contribuirá a facilitar la creación de bases de datos de inventarios de archivos y documentos completos.

Otro de los objetivos del proyecto cumplido ampliamente es demostrar la viabilidad del uso de XML para hacer accesible a través de Internet descripciones no normalizadas de archivos, bibliotecas y museos previamente existentes. De hecho, todas las descripciones y documentos incorporados al sistema COVAX han sufrido procesos de conversión y reelaboración.

COVAX ha supuesto, una aplicación concreta de XML y del concepto de interconexión entre sistemas que supone Z39.50. La normativa y protocolos para la consulta a bases de datos distribuidas debe mejorar y en esta dirección van las propuestas de superación del protocolo Z39.50 para adecuarse a HTTP e incluir XML[10]. Sin embargo, un modelo como el de COVAX ha mostrado su validez para la aplicación a cualquier entorno, no sólo el de archivos, bibliotecas y museos, sino también para la distribución de información sobre productos de aprendizaje electrónico (*e-learning*) o información turística.

Por último, como se ha mostrado en el proceso de validación del sistema, el principio de realizar búsquedas de documentos a través de diferentes dominios (archivos, bibliotecas y museos) del que partía COVAX, ha sido fuertemente refrendado por los usuarios finales. Independientemente de las mejoras que puedan introducirse en la interfaz del prototipo a la hora de la ordenación y presentación de los datos, en general los usuarios encuentran que esta posibilidad es de enorme interés para la investigación.

Un concepto aplicable

Covax no es un producto acabado, pero sí un concepto aplicable a resolver los problemas de las instituciones de memoria en su proceso de adaptación al entorno digital o, mejor, en su proceso de hibridación: su ampliación hacia instituciones que ofrecen colecciones digitales (junto a sus colecciones analógicas) y servicios digitales (junto a servicios analógicos). Es especialmente adecuado para la constitución de colecciones virtuales que reúnan los recursos ofrecidos por un número de instituciones con colecciones relacio-

[9] La simplificación de la definición de tipo de documento desarrollada para registros MARC, que fue publicada una vez finalizado el proyecto, ha dado lugar a MARCXML Schema, aún demasiado sujeta al formato MARC. Véase: Marc in XML http://www.loc.gov/marc/marcxml.html

[10] Véase ZING, Z39.50-International: Next Generation http://www.loc.gov/z3950/agency/zing/zing.html

nadas entre sí (aunque también puede ofrecer acceso a colecciones completamente diferentes, permitiendo en primera instancia al usuario buscar en y navegar por las descripciones de alto nivel de los centros en busca de las colecciones más adecuadas a su interés).

Pero el camino abierto por COVAX no se detiene aquí. Con el *Archivo virtual* o este proyecto (y otros proyectos de investigación e innovación en marcha), la Residencia se embarcó en lo que, visto en perspectiva, ha resultado ser un proceso global de adaptación a la sociedad de la información. Este proceso, según ha ido precisándose con el tiempo, ha implicado el diseño de un modelo de actuación para las instituciones culturales (principalmente, las instituciones de memoria, pero también otros centros culturales de calidad) en el entorno digital.

Este modelo parte de la constatación de una serie de condiciones para que las instituciones culturales sigan cumpliendo con éxito (potencialmente, con un impacto mayor que en el pasado) su misión en el entorno digital. La primera es la necesidad de que se adapten conceptual y tecnológicamente al funcionamiento en red (muchas de sus nuevas funciones en el mundo digital, las cumplirán no separada, sino conjuntamente). También, la exigencia de digitalizar el patrimonio y la producción cultural para múltiples propósitos: para su preservación, para la creación de colecciones virtuales, para la difusión de contenidos culturales por los canales propios de la sociedad de la información, para la reelaboración de estos contenidos con propósitos de investigación y educativos). En tercer lugar, la importancia de crear instrumentos para facilitar estos usos: útiles para los investigadores, herramientas para la construcción de aplicaciones educativas. En cuarto lugar, la conveniencia de la normalización (la adopción de estándares lo más universales posibles para hacer *interoperables*, es decir, capaces de funcionar conjuntamente, los sistemas y aplicaciones). Por último, la exigencia de crear entornos *amigables* y adaptados a las demandas de los usuarios (especialmente, investigadores, estudiantes y profesores, consumidores de cultura de calidad).

En el camino hacia la concreción de este modelo, COVAX ha representado un paso de gran utilidad, que ahora exige su puesta a prueba en la creación de un servicio operativo de acceso distribuido a bases de datos que contengan descripciones de documentos de tipología diversa, y que debe completarse con soluciones creativas al resto de las insuficiencias de los sistemas de acceso a los recursos de las instituciones de memoria que se señalaban al principio. Ese es el camino que ha emprendido la Residencia de Estudiantes.