Estructuración de la información mediante XML:

Bonifacio Martín Galán
Departamento de Biblioteconomía y Documentación
Universidad Carlos III de Madrid bmartin@bib.uc3m.es

David Rodríguez Mateos
Departamento de Biblioteconomía y Documentación
Universidad Carlos III de Madrid pirio@bib.uc3m.es

Resumen: XML surgió como el lenguaje de marcado de documentos que sustituiría a HTML en la Web. Ambos lenguajes son herederos de SGML, el lenguaje de marcado estándar para la descripción formal y de contenido de los documentos (en contraposición a los lenguajes de marcado orientados a la presentación).

HTML se impuso por su sencillez y espectacularidad, hasta el punto de que, por una parte, las compañías de software lo orientaron hacia formatos de presentación, y por otra, numerosas empresas apostaron por HTML para organizar documentos en entornos corporativos, con escaso éxito, por su reducida capacidad para estructurar documentos.

XML trató de ser la solución: una versión reducida de SGML con evidentes valores documentales (al igual que SGML), para definir estructuras formales y de contenido independientes de la presentación. No obstante, buen número de empresas han adoptado XML como herramienta de intercambio de datos, obviando este valor documental.

En el campo de la documentación, el interés de XML radica en su uso, tanto en Web como en sistemas corporativos, como herramienta de definición de estructuras y de descripción de contenidos. En ambos casos, estructura y/o descripción se incluyen dentro del documento y permiten la reutilización de partes del mismo.

No obstante, debe aún mantenerse la atención sobre HTML (que continuará como lenguaje de marcado para documentos con escasa estructuración) y sobre la incierta evolución de algunas aplicaciones de XML (XLink, Namespaces, XSL, etc.), de evidente interés para los documentalistas.

Palabras clave: Lenguajes de marcado de texto; SGML; HTML; XML; XHTML; Estructuración y organización de documentos electrónicos; Definición de tipo de documento (DTD); Esquemas XML; Gestión de información electrónica.

El lenguaje de marcado de documentos HTML (*Hypertext Markup Language*) participó en la popularización de la *World Wide Web*, como soporte de los documentos que la componen, impulsando además su uso en numerosas redes informáticas corporativas. Para poder comprender su pretendida evolución hacia otros len-

guajes de este tipo, como XML (*eXtensible Markup Language*)[1], debe observarse la evolución sufrida por este modelo de construcción de documentos electrónicos. Los grandes pilares de esta evolución se fundamentan en la importancia de los lenguajes de marcado de texto (como elemento estructural del modelo) así como en la búsqueda de un formato sencillo y universal de documento electrónico que haga factible el paradigma de efectividad en la gestión de información en entornos informatizados.

Al nivel más simple, todos los documentos impresos llevan un marcado o conjunto de reglas de codificación que les hace comprensibles (espacios en blanco entre las palabras, signos de puntuación, etc.). Un concepto más restrictivo de la marca proviene del mundo de la edición: los manuscritos eran anotados con instrucciones para el procesamiento por parte del cajista de la imprenta. Con la implantación de los procesadores de texto, el concepto de marca se hizo extensible a todos los códigos de marcado de formato que se insertaban en los documentos electrónicos[2]. Este marcado se orienta exclusivamente a la correcta presentación formal del documento electrónico.

No obstante, es posible utilizar el marcado de los documentos con fines documentales, al permitir separar elementos lógicos o estructuras del documento, o bien, realizar abstracciones del original para extrar la metainformación del mismo.

En ambos casos, el conjunto de las reglas que establecen qué tipo de marcas han de ser utilizadas, de qué modo se distinguirán las marcas del texto del documento y cómo se insertarán éstas (la gramática y su sintaxis), y cuáles son las marcas permitidas en cada una de las partes del texto, es lo que se conoce como un lenguaje de marcado[3] (*markup language*). Teniendo en cuenta la doble vertiente de la utilización de la marca en los textos electrónicos es posible distinguir sendos tipos de lenguaje de marcado:

— Lenguajes de marcado orientados a la presentación de los documentos (*procedural markup*): especifican cómo debe ser procesado el texto para su salida a través de diversos medios (pantalla de ordenador, impresora…)[4]. Este tipo de lenguajes cuenta con una serie de desventajas para la gestión

---

[1] Uno de los primeros estudios descriptivos sobre XML que se realizaron dentro de nuestro entorno profesional en nuestro país fue presentado en la anterior edición de FESABID por Fernando SANTAMARÍA GONZÁLEZ. «XML (*eXtensible Markup Language*): Nuevo estándar para la descripción de documentos en la World Wide Web». En: *Jornadas Españolas de Documentación* (6.ª Valencia. 1998). Valencia: FESABID, Associació Valenciana d'Especialistes en Información, 1998, pp. 819-825.

[2] Lou BURNARD. «Markup and Markup Languages [documento HTML]». En: *What is SGML and How Does it Help?* Oxford: The Humanities Computing Unit, Oxford University Computing Service, 1999. Disponible en http://www.hcu.ox.ac.uk/TEI/Papers/EDW25/W25C.htm (consultado el 5 de junio de 2000). Publicado originalmente en: *Computers and the Humanities*, v. 29, 1995, pp. 41-50.

[3] Pete JOHNSTON. *What you always wanted to know about SGML, HTML and XML* [documento HTML]. Edinburgh: University of Glasgow Archives and Business Records Centre, 1 December 1998. Disponible en http://www.gla.ac.uk/InfoStrat/socarcpj/ (consultado el 13 de junio de 2000).

[4] Algunos de los formatos de este tipo más conocidos son el RTF (*Rich Text Format*) de la compañía Microsoft, el PDF (*Portable Document Format*) de la compañía Adobe, y el PostScript.

documental: no registran la estructura lógica del documento, por lo que no aportan información de tipo semántico o estructural; son muy poco flexibles: cualquier cambio en la presentación del documento implica modificar su marcado; en su mayoría, son lenguajes específicos de un sistema de procesamiento propietario, lo cual reduce la «portabilidad» de dichos documentos.

— Lenguajes de marcado orientados a la descripción formal y de contenido de los documentos (*descriptive markup*), es decir, a la identificación de las piezas lógicas que conforman la estructura de un documento (para formalizar estructuras normalizadas de tipos documentales) y/o a la descripción del contenido informacional del mismo (preparando los documentos electrónicos para ser almacenados y recuperados). Son lenguajes más flexibles, diferenciando entre lo que el documento es y cómo debe ser presentado éste.

A finales de los 60, un equipo de investigadores de IBM encabezado por Charles F. Goldfarb desarrolló GML (*Generalized Markup Language*) para gestionar la documentación de valor legal de dicha compañía. Este es el primer lenguaje no propietario de marcado de texto (independiente del sistema en el se crean los documentos y de la plataforma en la que circulan) capaz de definir las estructuras lógicas de cualquier tipo de documento basándose en una serie de normas restrictivas[5].

El principal lenguaje en este campo lo constituirá SGML (*Standard Generalized Markup Language*), convertido en 1986 en norma internacional (ISO 8879:1986)[6]. La importancia de este metalenguaje[7] radica, para los profesionales de la documentación automatizada, en aspectos como el principio de independencia de los datos para una gestión de la información abierta[8], un modelo de creación de gramáticas y sintaxis para el marcado del contenido de los documentos a través de etiquetas definidas por el usuario, la presentación de documentos a través de hojas de estilo, la representación de las relaciones hipertextuales explícitas e implícitas en y entre documentos[9], y la descripción normalizada de la estructura lógica de tipos documentales.

---

[5] Véase Liora ALSCHULER. *ABCD… SGML: A User's Guide to Structured Information*. London [etc.]: International Thomson Computer Press, 1995, pp. 5-6.

[6] Una buena descripción técnica de SGML se encuentra en la obra de Charles F. GOLDFARB. *The SGML Handbook*. Oxford: Clarendon Press, 1990; o en el texto de Neil BRADLEY. *The Concise SGML Companion*. Reading (Massachusetts) [etc.]: Addison Wesley, 1997.

[7] SGML no define en sí cómo ha de ser un determinado documento y cuál ha de ser su estructura sino que posibilita al usuario definir lenguajes de marcado propios, conocidos como «aplicaciones SGML», para ser aplicados a cualquier tipo de documento electrónico.

[8] La idea central del OIM (*Open Information Management*) reside en que la información debe poder ser procesada por cualquier programa, independientemente del programa que generó en inicio esos datos.

[9] Estas normas, que se desarrollaron de forma paralela a la evolución de SGML, fueron, para el caso de la definición de la presentación de documentos, DSSSL (*Document Style Semantics and Specification Language*) y, para la definición de relaciones hipertextuales entre documentos, HyTime.

SGML estructura los contenidos informativos del documento de forma similar a una base de datos, pero manteniendo la integridad documental y, al tiempo, la independencia de los datos del documento (éstos pueden ser reutilizados para generar, a su vez, nuevos documentos). Este lenguaje permite definir, para la mayor parte de los documentos, su agrupación en tipos que mantienen una estructura lógica identificable, es decir, siguen un patrón común en la inclusión de piezas o elementos (datos) que conforman el corpus de dicho documento. En palabras del propio Goldfarb, que ponía como ejemplo patrones de documentos de tipo jurídico, «cada uno de ellos es un *tipo de documento* y la definición formal que describe cada tipo se denomina *definición de tipo de documento* (DTD)»[10]. Por tanto, la DTD, obligatoria en SGML, especifica qué elementos conforman la estructura de un tipo de documento, qué lugar ocupan dentro del modelo jerarquizado de dicha estructura, y cómo se relacionan entre sí.

Son innumerables las instituciones que adoptaron SGML como modelo de estructuración de contenidos documentales. Algunos ejemplos[11]: la *Air Transport Association* (ATA), generadora de un gran número de DTDs para documentación técnica en el mundo de la aviación; la *International Press Telecommunications Council* y la *Newspaper Association of America*, que desarrollan DTDs para los servicios de prensa electrónica, desde la primera UTF (*Universal Text Format*) hasta el actual NewsML; o, la ampliamente utilizada DTD DocBook, para la descripción de manuales técnicos, principalmente de hardware y software, mantenida por la *Organization for the Advancement of Structured Information Standards* (OASIS)[12].

Pero la aplicación SGML que mayor repercusión ha tenido es HTML (*HyperText Markup Language*) En 1989 Tim Berners-Lee, investigador del CERN[13], basándose en la sintaxis de SGML, desarrolló este modelo de documento electrónico sencillo para compartir y gestionar información electrónica estructurada y normalizada en un entorno corporativo pero, a diferencia de SGML, para un uso por parte de personal no especialista en documentación: primaba la filosofía del hipertexto para crear vínculos entre los documentos de dicha institución y otras instituciones colaboradoras. El éxito fue de tal magnitud que Berners-Lee denominó a este sistema la *World Wide Web*.

En origen, HTML carecía de una definición de tipo de documento (DTD): se trataba únicamente de un sencillo lenguaje de marcado de texto que establecía una colección fija de marcas de elementos generalizados y descriptivos, aplicados en

---

[10] Charles F. GOLDFARB, Paul PRESCOD. *Manual de XML*. Madrid [etc.]: Prentice Hall Iberia, 1999, p. 11.

[11] Robin Cover mantiene desde mediados de la década pasada un magnífico recurso en la Web sobre los principales desarrollos que han ido surgiendo entorno a SGML. Véase en http://www.oasis-open.org/cover/sgml-xml.html

[12] Para una mayor información sobre esta aplicación SGML y la poderosa organización que lo sustenta véase http://www.oasis-open.org/docbook/

[13] Actual *Organisation Européenne pour la Recherche Nucléaire*.

documentos hipertextuales. La rápida popularización de la Web, al convertirse en un vistoso modelo de acceso a la información a través de Internet, hizo necesaria la creación del *World Wide Web Consortium* (W3C), cuyos fines incluían crear y normalizar una DTD para los documentos HTML, la cual solo se alcanza a partir de la versión 3.2 de HTML[14].

SGML quedó circunscrito a la publicación, gestión e intercambio de documentos electrónicos en grandes instituciones, dados los altos costes que requieren los complejos sistemas de información basados en este metalenguaje. HTML iba destinado al gran público por su sencillez, espectacularidad y grandes perspectivas de negocio vislumbradas por las principales compañías informáticas: en gran medida, el gran público está acostumbrado a primar la apariencia visual, creando documentos electrónicos a través de procesadores de texto, sin entender la importancia de marcar las la estructura lógica de un documento electrónico. Grandes empresas de software, con Netscape y Microsoft a la cabeza, presionaron hasta obligar al W3C a introducir, junto a las etiquetas HTML iniciales que marcaban piezas estructurales básicas, etiquetas orientadas a la presentación, obligando al Consorcio a que fueran aceptadas dentro de la especificación oficial. Las consecuencias empezaron a ser fatídicas para la interoperatibilidad de la Web.

El mundo de la documentación automatizada participó en este *boom*: además de crear documentos hipertextuales para la Web, utilizaron HTML como modelo para la gestión de documentos electrónicos en entornos corporativos informatizados, debido en parte a la aguda visión comercial de las principales compañías informáticas. Así surgen las «Intranets», que tanta literatura científica han generado en los últimos años[15]. Lamentablemente, este modelo de marcado hipertextual con HTML se reveló insuficiente para cumplir con varias premisas básicas: la estabilidad de un modelo de descripción de documentos electrónicos, la estructuración de la información electrónica para su almacenamiento en bases de datos, así como la potencialidad del formato para la recuperación exacta de información.

XML (*eXtensible Markup Language*) fue la respuesta aportada por el W3C en 1998, con la publicación de la versión 1.0[16] como recomendación de 10 de febrero de dicho año, aún vigente. La recomendación[17] establece que «el lenguaje extensible de marcas (XML) es un subconjunto de SGML (…) Su objetivo es permitir que

---

[14] Un buen análisis sobre la evolución sufrida de forma paralela por la WWW y el lenguaje HTML se encuentra en J. Tomás NOGALES FLORES. «La Revolución de la World Wide Web». En: Mercedes CARIDAD SEBASTIÁN (coord.). *La Sociedad de la Información: Política, Tecnología e Industria de los contenidos*. Madrid: Centro de Estudios Ramón Areces, 1999, pp. 175-212.

[15] Destacamos la obra de Joan BANNAN. *Intranet Document Management*. Reading, Massachusetts: Addison-Wesley, 1997. Asimismo, uno de los autores de esta comunicación, Bonifacio MARTÍN, defendió, al finalizar sus cursos de doctorado en la Universidad Carlos III de Madrid, la tesina, *La Intranet Documental y la Gestión de la Información en Entornos Corporativos* (21 de septiembre de 1998).

[16] La recomendación del W3C sobre XML 1.0 se encuentra accesible en la dirección http://www.w3.org/TR/1998/REC-xml-19980210

[17] Utilizamos la versión «oficial» en español, referida desde el propio W3C, accesible en la dirección http://wwwest.uniandes.edu.co/~l-arcini/Spec.html.

SGML genérico pueda ser servido, recibido y procesado en la web en la misma manera que hoy es posible con HTML. XML ha sido diseñado de tal manera que sea fácil de implementar y buscando interoperabilidad tanto con SGML como con HTML.»

Se trata, pues, de una forma simplificada de SGML[18], adecuada para operar en la Web: un metalenguaje que permite definir al usuario lenguajes propios de marcado de documentos electrónicos en un entorno informático con tecnología Web. Pero, ¿por qué no se decidió utilizar directamente SGML? La idea básica que se ha difundido afirma que SGML es apto para trabajar con grandes volúmenes de información con representaciones estructurales muy complejas, pero demasiado complicado en un entorno no dominado por especialistas en información y documentación, además de haberse construido sin pensar en las actuales aplicaciones en línea[19]. Aunque todo documento XML es documento SGML por definición, se establecía que SGML quedaba para la creación y almacenamiento de documentos y datos en entornos corporativos y que XML sería utilizado para el suministro de documentos en la Web.

¿Dónde queda entonces HTML? Para el W3C, las posibilidades de este lenguaje se veían agotadas tras HTML 4.0, por lo que se trabajó en la transformación del mismo para adaptarlo a la sintaxis de XML. El resultado fue XHTML (*eXtensible HyperText Markup Language*), cuya versión actual, la 1.0, se rige por la recomendación del W3C de 26 de enero de 2000[20]. Los tradicionales documentos hipertextuales existentes en Internet serían ahora un determinado tipo de documento definido bajo la sintaxis XML, con una DTD mucho más restrictiva que la existente para HTML y, además, «extensible»: puede incorporar otros elementos, y sus correspondientes etiquetas, sin estar definidos en la DTD, siempre y cuando estén construidas siguiendo la sintaxis XML[21]. Todo ello no significa el fin de HTML,

---

[18] La recomendación XML (26 páginas) es una profunda simplificación del estándar SGML (más de 500 páginas). Véase el *white paper XML for Managers: Evaluating SGML vs. XML from a Manager's Perspective* en http://www.arbortext.com/Think_Tank/XML_Resources/XML_for_Managers/xml_for_managers.html

También puede consultarse la comunicación de Antonio DE LA ROSA PIÑERO. «Entornos documentales WWW: entorno XML». En: *Jornadas Andaluzas de Documentación* (2.ª Granada. 1999). Granada: Asociación Andaluza de Documentalistas, 1999, pp. 357-372.

[19] Thomas A. POWELL. *HTML: Manual de referencia*. Madrid [etc.]: McGraw-Hill Interamericana de España, 1998, p. 610. Resultan de gran interés todas las reflexiones que el autor realiza en el capítulo «XML: Mas allá de HTML», pp. 609-631.

No obstante, el comité de trabajo auspiciado en 1996 por el W3C, conocido por el nombre de *Comité de Revisión Editorial de SGML*, cambió de filosofía y de nombre a *The XML Working Group*, integrado por compañías de la talla de Microsoft, Netscape, Sun, etc. En el fondo de esta polémica subyace una lucha de poder para establecer quién y de qué modo se deben establecer los estándares que guíen la evolución de la Web. Una interesante reflexión se realiza en Liora ALSCHULER. *Setting the Standard* [documento HTML]. XML.com, 10 de mayo de 2000. Disponible en http://www.xml.com/pub/2000/05/10/standards/index.html (consultado el 5 de julio de 2000).

[20] Véase en http://www.w3.org/TR/xhtml1/

[21] Para una mayor comprensión inicial sobre XHTML y las diferencias respecto a HTML recomendamos la lectura de Peter WIGGIN. *XHTML: The Clean Code Solution* [documento HTML]. The O'Reilly Network, 24 de abril de 2000. Disponible en http://www.oreillynet.com/pub/a/network/2000/04/28/feature/xhtml_rev.html (consultado el 16 de junio de 2000).

como algunos investigadores han señalado. Al contrario, el lenguaje HTML se ha convertido recientemente en un estándar internacional, la ISO/IEC 15445:2000, por lo que su vida se promete duradera[22].

Para entender las aportaciones de XML a la organización de información electrónica es necesario señalar que, de forma general, XML ha sido orientado hacia dos tipos de aplicaciones[23]: a) aquellas en las que se contempla al documento como un todo (*document applications*), concentrándose en la definición de la estructura documental para la gestión electrónica de los documentos, independientemente de su presentación; b) aquellas en las que el elemento fundamental es la gestión, almacenamiento e intercambio de los datos (*data applications)*, por parte de sistemas de bases de datos. Las grandes compañías de software han apostado por esta opción, como modelo para la gestión del comercio electrónico a través del medio Internet, aunque tal vez XML esté todavía lejos de ser el elemento integrante de todas estas tecnologías[24].

Para los profesionales de la documentación automatizada, el mundo de la información electrónica integra ambas visiones. Esta división entre datos reales y documentos (entendiendo éstos últimos solo como fuentes de las que se extraen los datos, o bien, sólo como productos generados a partir de los datos), no existe de forma tan taxativa. En palabras del propio Goldfarb, «los documentos son una fuente de información para la humanidad mucho más significativa, y capaces de almacenar estructuras de información mucho más ricas que las bases de datos.»[25] Por tanto, el interés de XML para los documentalistas consiste en representar, a través de este metalenguaje, estructuras lógicas de tipos documentales, y además, aislar los datos significativos capaces de caracterizar una correcta descripción del contenido del documento para su almacenamiento y recuperación, con independencia del medio en el que los documentos electrónicos serán suministrados o presentados.

Aunque estructura y contenido van íntimamente ligados, existen enfoques diferentes en el tratamiento de dichas materias por investigadores e instituciones implicadas en el desarrollo de estándares de aplicación en Internet o en espacios corporativos electrónicos. Por un lado, han surgido diferentes modelos de descripción y organización de los contenidos informativos existente en el espacio Web:

---

[22] Mientras ISO vio a HTML como estándar de facto (ISO HTML), el W3C estaba ya trabajando en la transformación del HTML al entorno XML (XHTML). De hecho, HTML es ya un estándar internacional de la ISO y la ICE (*the International Electrotechnical Commission*) desde mayo de 2000, conocido como ISO/IEC 15445:2000 (E) (existe una versión de acceso público en http://woodworm.cs.uml.edu/~rprice/15445/15445.html). Véase Todd FRETER. *XML: It's the Future of HTML* [documento HTML] Sun Microsystems, 2 de junio de 1998. Disponible en http://www.sun.com/980602/xml/ (consultado el 9 de diciembre de 1999).

[23] Benoît MARCHAL. *XML by Example*. Indianapolis: Que, 2000, p. 28.

[24] Peter FISCHER. «XML is Not Yet A Cornerstone Technology». *Application Development Trends*, v. 7, n.º 4, April 2000, pp. 55-60. Disponible en formato HTML en http://www.adtmag.com/Pub/apr2000/fe401a.cfm (consultado el 6 de junio de 2000).

[25] Charles F. GOLDFARB, Paul PRESCOD., *op. cit*., p. v.

los llamados metadatos[26], o información sobre los datos. Por otro lado, existe otra forma más directa de incluir la metainformación dentro de los propios documentos XML, utilizando para ello las capacidades que tiene los elementos XML de autodescribir metadatos[27].

En el marco de esta comunicación, la piedra angular del modelo de documentos XML es la definición de estructuras lógicas de tipos documentales: todos los documentos son susceptibles de ser divididos en piezas lógicas, o «elementos», que caracterizan las diversas estructuras lógicas de tipos documentales diversos. Los elementos pueden incluir otros elementos, información más específica declarada en los «atributos», así como frases y palabras, que constituyen el texto propiamente dicho. Para ello resulta fundamental conocer las piezas informativas del documento, cómo se relacionan unas con otras y, en menor medida, cuáles son los items físicos que constituyen la información[28]. Todo ello se puede formalizar en XML a través de una Definición de Tipo de Documento (DTD) o, más recientemente, a través de los llamados *XML Schemas*.

El concepto de DTD está relacionado con la distinción de clases de documentos que se establecen en XML: documentos bien formados y documentos válidos. Un documento **bien formado** se ajusta a la sintaxis XML pero carece de DTD: tiene una estructura marcada por el etiquetado de las piezas lógicas siguiendo un modelo, pero dicha estructura es única para ese documento. Un documento **válido** es un documento estricto: se construye siguiendo las reglas de la sintaxis XML, y sus elementos siguen una estructura y reglas de relaciones que se ajustan a lo establecido para un determinado tipo de documento, descrito en una DTD (o en un esquema). Esta dualidad se entiende dado que XML es un SGML abreviado y operativo en la Web: no es tan importante que los documentos XML se ajusten a

---

[26] Puede seguirse la actividad del W3C sobre los metadatos en la página web http://www.w3.org/Metadat/Activity.html que mantiene Ralph Swick. Recomendamos además la lectura de Jian QUIN. «Representation and Organization of Information in the Web Space: From MARC to XML». *Informing Science*, v. 3, n.º 2, 2000, pp. 83-88. Disponible en formato PDF en http://inform.un/Articles/Vol3/v3n2p83-88.pdf (consultado el 4 de julio de 2000).

En especial, sobre el último desarrollo en este campo, el RDF (*Resource Description Framework*), recomendamos la consulta de Eva M.ª MÉNDEZ RODRÍGUEZ. «RDF: Un modelo de metadatos flexible para las bibliotecas digitales del próximo milenio». En: *Jornades Catalanes de Documentació (7.ª Barcelona. 1999)*. Barcelona: Col.legi Oficial de Bibliotecaris-Documentalistes de Catalunya, 1999, pp. 487-498. Disponible también en: http://www.bib.uc3m.es/~mendez/publicaciones/7jc99/rdf.htm.

[27] Aspecto éste de vital importancia para la gestión documental en entornos corporativos, tal y como lo expresa Todd Freter, de la compañía Sun Microsystems, al decir: «When the elements carry self-describing metadata with them, systems that understand XML syntax can operate on those elements in useful ways, just like a traditional document management system can.» Véase en Todd FRETER. *XML: Document and Information Management* [documento HTML]. Sun Microsystems, 8 de septiembre de 1998. Disponible en http://www.sun.com/980908/xml/ (consultado el 9 de diciembre de 1999).

[28] La definición de la estructura física de los documentos XML viene determinada por las «entidades», que permiten fragmentar dichos documentos en múltiples piezas, pero sin olvidar que forman un todo coherente. Las entidades permiten hacer realidad una de las mayores ventajas que aportan estos modelos: la reutilización de piezas de información de los documentos para formar nuevos documentos.

tipos documentales concretos pues, en muchos casos, o bien constituirán un nuevo modelo para insertar contenidos interactivos en este espacio (con una funcionalidad parecida al HTML pero con mayores capacidades), o bien, como se está utilizando hoy en día por parte de muchas empresas, dichos documentos XML no serán «visibles» para los usuarios de la red, sino que constituirán un modelo de intercambio de datos entre los nuevos sistemas de bases de datos orientados al comercio electrónico.

Pero, dado que los orígenes y filosofía de XML están en SGML (es decir, en la gestión de la información electrónica en espacios corporativos), será preciso normalizar a través de DTDs las estructuras de los documentos que conforman la tipología documental empleada en una organización para la correcta gestión de los mismos. Esta necesidad se apoya en tres importantes razones[29]: 1) previene posibles luchas de etiquetado (*tag wars*) al establecer un modelo normalizado para cada uno de los tipos documentales de la organización, aceptado por los miembros de la misma; 2) la creación de una especificación formal de tipo de documento, aunque larga y compleja, proporciona mayor eficacia a las organizaciones al asegurar la integridad y veracidad de los datos tratados en los sistemas de gestión de documentos e intranets corporativas; 3) la DTD resulta válida para definir la metainformación (metadatos) de importancia para la organización, expresada en el lenguaje que dicha organización establezca, e imbuida en los elementos y, principalmente, en los atributos de los mismos.

Si las DTDs son consideradas como piezas esenciales para la estandarización documental dentro de las organizaciones, muchos especialistas de la documentación deberán reorientarse dentro de sus instituciones de trabajo hacia el conocimiento y uso de estas técnicas. En muchos casos será posible utilizar directamente DTDs ya creadas[30], estandarizadas por algunas organizaciones relacionadas con la tipología documental propia de nuestra institución. En otros casos, este consenso no será factible debido a la generalidad de las normas creadas o a su falta de adecuación a las necesidades estructurales de la documentación de una institución en particular. Será preciso entonces, partiendo de experiencias ajenas, que los documentalistas aprendan a definir y desarrollar DTDs específicas para aquellos tipos documentales de sus entornos documentales corporativos[31], y que tales DTD puedan ser normalizadas en el futuro por la respectiva comunidad científica.

---

[29] Chet ENSIGN. «Structure rules!: Why DTDs matter after all». *Markup Languages: Theory and Practice*, v. 1, n.º 1, winter 1999, p. 104. Disponible en formato PDF en http://mitpress.mit.edu/journals/MLANG/ensign.pdf (consultado el 24 de de enero de 2000).

[30] Normalmente, DTDs SGML adaptadas a la simplificación de las DTDs XML. La mejor fuente de información en Internet para consultar desarrollos normalizados de vocabularios XML lo constituye el sitio Schema.net, de James Tauber, en http://www.schema.net/

[31] Como señala acertadamente Natanya Pitts, «XML pone un gran énfasis en los estándares, y la mayoría de las DTDs XML que hoy en día se están desarrollando se encuentran bastante bien documentadas, por lo que el proceso de aprender de otros sí es un modo fiable de convertirse en un maestro del diseño de DTDs.» En: Natanya PITTS. *XML*. Madrid: Anaya Multimedia, 1999, p. 221.

Existe otra alternativa[32] al modelo de DTD: los citados «esquemas»[33]. La idea principal del modelo de esquemas consiste en que, en XML, la estructura lógica de muchos documentos electrónicos se asemeja a la estructura que puede tener una base de datos, tanto relacional como orientada a objetos[34]. Este modelo está influenciado por los informáticos desarrolladores de las actuales bases de datos, así como por las empresas informáticas con intereses en el desarrollo del comercio electrónico en Internet).

Además de esta diferenciación de base, existen otras características que hacen de los *XML Schemas* un modelo más potente para la modelización y definición de las estructuras válidas de los documentos XML: a diferencia de la DTD, que utiliza una sintaxis propia, los esquemas utilizan la misma sintaxis de los documentos XML; permiten definir una mayor riqueza y complejidad a las estructuras internas; permiten hacer uso de otros estándares XML acompañantes, como XLink, XPointer, XSL Namespaces, etc., para ampliar así sus capacidades de definición y actuación…

Existen diversas propuestas de esquemas XML. La primera de ellas, conocida por el nombre de *XML-Data*, fue presentada al W3C por un grupo de compañías con Microsoft a la cabeza, y lanzada como «nota» para su estudio y debate en enero de 1998[35]. Microsoft, junto a otras empresas y organizaciones, ha venido trabajando en la evolución de este modelo dando lugar al marco de trabajo BizTalk[36], que tanta polémica ha levantado en algunos sectores de la comunidad científica por la influyente presión hacia la adopción de modelos «normalizados» procedentes de las grandes compañías de software[37]. Junto a otros modelos de esquemas, destaca la propuesta del propio W3C, conocida por el nombre genérico de *XML Schema*, y definida en tres partes, aún en fase de borrador de trabajo (el último *working draft* es de 7 de abril de 2000)[38]. Se espera que esta propuesta imponga un poco de orden en este agitado entorno de intereses.

A modo de conclusión, los profesionales de la información y la documentación automatizadas nos encontramos frente a un reto profesional. Un nuevo lenguaje

---

[32] Simon ST. LAURENT. *Describing Your Data: DTDs and XML Schemas* [documento HTML]. XML.com, 1 de diciembre de 1999. Disponible en http://www.xml.com/pub/1999/12/dtd/index.html (consultado el 9 de diciembre de 1999).

[33] El concepto de «esquema» procede del mundo de las bases de datos y es empleado para denominar la serie de restricciones que se aplican a la estructura de una base de datos. Por extensión se aplicará también a los modelos que definen estructuras en tipos documentales XML.

[34] Yasser SHOHOUD. «XML's Grand Schema». *XML Magazine*, v. 1, n.º 3, summer 2000, pp. 38-43. Disponible en formato HTML en http://www.xlm.com/pub/1999/12/dtd/index.html (consultado el 5 de julio de 2000).

[35] Andrew LAYMAN… [*et al*.] *XML-Data: W3C Note 05 Jan 1998* [documento HTML]. World Wide Web Consortium, enero de 1998. Disponible en http://www.w3.org/TR/1998/NOTE-XML-data-0105/. Este esquema ofrecía unos tipos de datos inspirados en los tipos del lenguaje SQL.

[36] Véase el sitio Web en http://www.biztalk.org/

[37] Sin entrar en dicho debate, sí recomendamos, no obstante, la lectura de John TASCHEK. «The basic failure of XML is its premise». *PC Week*, v. 17, n.º 17, 24 de abril de 2000, p. 61. Disponible en formato HTML en http://www.zdnet.com/printerfriendly/0,6061,2551691-54,00.html (consultado el 5 de julio de 2000).

[38] Puede verse la información del W3C sobre este tema en http://www.w3.org/XML/Schema.

generalizado de marcado de texto está irrumpiendo con gran fuerza y debemos estar preparados para analizar con prudencia sus beneficios y sus contrapartidas. La amplitud de miras con la que fue confeccionado XML es una de sus grandes ventajas ya que permite que este modelo pueda ser aplicado al entorno documental de la Web así como a los espacios estancos de las redes informáticas corporativas.

En este contexto, XML cuenta con la dilatada experiencia de su «hermano mayor», SGML. Será, por tanto, un factor clave de aplicación de XML que los documentalistas aprendan a estructurar los contenidos informativos de los documentos electrónicos que se gestionan en sus organizaciones (contamos para ello con un *background* más que suficiente) así como a definir dichas estructuras a través de los modelos de normalización de tipos documentales planteados. No obstante, el entorno normativo de XML, así como el software capaz de procesar e interpretar correctamente estos documentos[39], se encuentran aún en fase de desarrollo, por lo que se hace prudente esperar un cierto tiempo para ver cómo y de qué manera evoluciona este modelo. Ello no es óbice para no entrar en la investigación y aplicación de este metalenguaje.

En cuanto a la aplicación de XML en la Web, el camino es más complejo: muchos usuarios seguirán llenando este espacio con documentos HTML, más ahora que este lenguaje de hipertexto se ha convertido en un estándar internacional. Por otro lado, las principales compañías informáticas con intereses en este campo están forzando la evolución de XML hacia una orientación dirigida al control de los datos[40]. Tal vez exista una transición hacia el modelo XHTML, debida a los esfuerzos del W3C, pero poco más puede esperarse de él. Para los profesionales de la documentación, el caballo de batalla será la organización y descripción de los contenidos informativos existentes en la Web, utilizando para ello técnicas y lenguajes derivados de XML, como RDF.

Mientras esta transición hacia XML se produce en los diferentes espacios electrónicos de actuación, las instituciones profesionales, académicas y de investigación de nuestro entorno profesional deberán implicarse directamente en el estudio, desarrollo evolutivo y aplicación de este lenguaje a la información y documentación electrónica de diversos tipos, características y procedencias, como de hecho ya se está produciendo.

---

[39] Por ejemplo, en el caso de los navegadores XML, queda aún tiempo para que éstos sean capaces de interpretar correctamente todas las piezas que componen el complejo puzzle de los documentos XML. Una comparativa sobre sus capacidades actuales se encuentra en Simon ST. LAURENT. *Browser XML Display Support Chart* [documento HTML]. XML.com, 2 de mayo de 2000. Disponible en http://www.xml.com/print/2000/05/03/browserchart/index.html (consultado el 15 de junio de 2000)

[40] Desde determinados sectores se lanzan propuestas y orientaciones sobre XML, tendentes a simplificar sus inclinaciones «documentales» originarias. Para una más amplia información sobre el tema véase Robin COVER. *Minimal XML* [documento HTML]. OASIS, 13 de abril de 2000. Disponible en http://www.oasis-open.org/cover/minimalXLM.html (consultado el 5 de julio de 2000).